Epidemics 17 (2016) 1–9

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Research Paper

Clustering of contacts relevant to the spread of infectious disease

Xiong Xiao[a,d], Albert Jan van Hoek[a], Michael G. Kenward[a], Alessia Melegaro[b], Mark Jit[a,c,*]

[a] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom
[b] DONDENA Centre for Research on Social Dynamics & Public Policy, Università Bocconi, Via Guglielmo Röntgen n. 1, 20136 Milan, Italy
[c] Modelling and Economics Unit, Public Health England, 61 Colindale Avenue, London NW9 5EQ, United Kingdom
[d] Department of Epidemiology and Biostatistics, West China School of Public Health, Sichuan University, Chengdu, China

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form 4 August 2016
Accepted 23 August 2016
Available online 26 August 2016

Keywords:
Clustering
Contacts
Infectious diseases
Mathematical modelling
Varicella-zoster virus

ABSTRACT

Objective: Infectious disease spread depends on contact rates between infectious and susceptible individuals. Transmission models are commonly informed using empirically collected contact data, but the relevance of different contact types to transmission is still not well understood. Some studies select contacts based on a single characteristic such as proximity (physical/non-physical), location, duration or frequency. This study aimed to explore whether clusters of contacts similar to each other across multiple characteristics could better explain disease transmission.
Methods: Individual contact data from the POLYMOD survey in Poland, Great Britain, Belgium, Finland and Italy were grouped into clusters by the *k* medoids clustering algorithm with a Manhattan distance metric to stratify contacts using all four characteristics. Contact clusters were then used to fit a transmission model to sero-epidemiological data for varicella-zoster virus (VZV) in each country.
Results and discussion: Across the five countries, 9–15 clusters were found to optimise both quality of clustering (measured using average silhouette width) and quality of fit (measured using several information criteria). Of these, 2–3 clusters were most relevant to VZV transmission, characterised by (i) 1–2 clusters of age-assortative contacts in schools, (ii) a cluster of less age-assortative contacts in non-school settings. Quality of fit was similar to using contacts stratified by a single characteristic, providing validation that single stratifications are appropriate. However, using clustering to stratify contacts using multiple characteristics provided insight into the structures underlying infection transmission, particularly the role of age-assortative contacts, involving school age children, for VZV transmission between households.

1. Introduction

Mathematical models of infectious disease transmission require assumptions about mixing between different subgroups in a population that can potentially lead to transmission between infected and susceptible individuals. The simplest assumption is that everyone has the same probability of contacting each other, but this can sometimes lead to misleading results (Keeling and Rohani, 2008). Indeed, many infection control interventions such as vaccinating children (Thorrington et al., 2015) or closing schools during a pandemic (House et al., 2011) are predicated on the assumption that certain subgroups in the population are the main transmitters.

A more realistic assumption is to subdivide the population based on some characteristic, and introduce a matrix of contact rates capable of transmitting infection between each subgroup, called the "who acquires infection from whom" (WAIFW) matrix (Vynnycky and White, 2010). Age is the characteristic most commonly used as a source of heterogeneity in mixing patterns. A model with age-stratified contact rates can be fitted to age-specific data on infection history (such as sero-epidemiological data, which marks the prevalence of previous infection) to estimate the age-specific effective contact rates in the WAIFW matrix.

To inform the elements of the WAIFW matrix, the number of social contacts that individuals in different age groups report can

Abbreviations: AIC, Akaike Information Criterion; AICc, small-sample-size corrected Akaike Information Criterion; ASW, average silhouette width; BIC, Bayesian Information Criterion; PAM, partitioning around medoids; VZV, varicella-zoster virus; WAIFW, who acquires infection from whom.

* Corresponding author at: Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom.
E-mail addresses: Xiong.Xiao@lshtm.ac.uk (X. Xiao), Albert.VanHoek@lshtm.ac.uk (A.J. van Hoek), Mike.Kenward@lshtm.ac.uk (M.G. Kenward), alessia.melegaro@unibocconi.it (A. Melegaro), mark.jit@lshtm.ac.uk (M. Jit).

http://dx.doi.org/10.1016/j.epidem.2016.08.001

be empirically measured and used as proxies to the actual contact rates underlying transmission (Beutels et al., 2006; Edmunds et al., 1997; Wallinga et al., 2006). The largest such study is a diary-based survey of 7290 participants in eight European countries collected in 2006 as part of the POLYMOD project (Mossong et al., 2008). Since then contact studies have been carried out in other parts of the world using similar methodology.

In these studies, participants are asked to record the contacts that they have made over a single day and classify them using a number of characteristics (such as physical/non-physical, long/short, home/school/work etc.) Since it is unrealistic to measure effective contacts (i.e. contacts that can transmit a particular infection) between individuals directly, some form of self-reported social behaviour such as face-to-face conversation or skin-to-skin contact is used as a proxy. It is assumed that the age distribution of these social contacts is related to the age distribution of effective contacts by a constant proportionality factor, an assumption referred to as the "social contact hypothesis" (Wallinga et al., 2006). Hence the age-specific transmission matrix is completely described by estimating this factor. Subsequent analyses (Melegaro et al., 2011) found that stratifying the WAIFW matrix according to different characteristics of contacts (with a different proportionality constant for each type of contact) significantly improved the goodness of fit of the models to serological markers of past infection for respiratory infections. This implies that different characteristics of social contact contribute differently to infection transmission.

A previous study explored the use of formal clustering algorithms to group POLYMOD survey respondents based on the number and location of their contacts (Kretzschmar and Mikolajczyk, 2009). The study found that respondents across different countries fell into a similar range of contact profiles, with the work, school and household contact profiles most common. However, this does not tell us whether there are certain types of contacts (rather than respondents) which may be particularly relevant for the transmission of particular infectious diseases. Furthermore, the relevant clusters of contacts may be disease-specific, since different infections have separate routes of transmission.

Hence an alternative approach would be to explore what kind of contacts (rather than respondents) are most relevant to infection transmission. The only previous work in this area has focused on single dimensions of contacts (Melegaro et al., 2011). This simply indicated that "intimate" (i.e. physical, home, long-duration and frequent) contacts are better able to explain age-dependent patterns in the acquisition of serological markers for varicella-zoster virus (VZV) and parvovirus B19, the two infection examined in the study. Since the dimensions of intimacy are highly correlated, it may be more informative to take all the characteristics of social contacts into account collectively when stratifying the WAIFW matrix.

In particular contacts made by different respondents could be grouped into clusters, and then examined to identify which clusters best explain patterns of infection acquisition in the corresponding populations. Here, we explore the use of clustering algorithms to determine clusters of social contacts which are similar to each other based on multi-dimensional characteristics of social contacts.

A range of clustering algorithms have been developed such as hierarchical clustering, partitioning clustering and latent class clustering (Everitt et al., 2011). In hierarchical clustering, each element is plotted on a graph to determine the optimal number of clusters. When the number of elements becomes very large (such as the thousands of contacts for each country in the POLYMOD survey), this graphical method becomes very cumbersome. In latent class clustering, a multivariate distribution is imposed on the data and the validation of the clustering results depends on several assumptions such as the parametric form of the multivariate distribution, the dependency between variables and the approximation of likelihood estimation. Since social contact data include different types of variables (binary and ordinal), some of which are structured, it is difficult to identify an appropriate multivariate distribution to describe the data. Hence we used partitioning clustering, and in particular the *k* medoids method to classify contacts.

We then investigate whether these clusters enhance our understanding of transmission patterns by using them to fit a transmission dynamic model to age-dependent patterns in the acquisition of varicella-zoster virus serological markers, as an example of a childhood respiratory infection with clear, long-lived markers of past infection and no vaccination history at the time of data collection.

2. Methods

2.1. Data sources

Age-specific contact matrices were constructed using social contact data from participants of the POLYMOD project (Mossong et al., 2008) living in Poland (15808 contacts, 1003 participants), Great Britain (11052 contacts, 996 participants), Belgium (8810 contacts, 747 participants), Finland (10319 contacts, 973 participants) and Italy (15788 contacts, 842 participants). Contacts with any missing information were discarded, so our dataset differs slightly from previous analyses (Melegaro et al., 2011). Contact matrices were adjusted for population size and reciprocity using well-described procedures (Melegaro et al., 2011; Wallinga et al., 2006).

Models were fitted to data on the presence of antibodies to VZV from serum samples collected in 1996 from 1300 participants aged 0–19 in Poland, 2091 participants aged 0–20 in England and Wales (fitted to Great British contacts), 2760 participants aged 0–39 in Belgium, 2500 participants aged 0–79 in Finland and 2517 participants aged 0–79 in Italy (Vyse et al., 2004). The samples were collected from unlinked anonymised (apart from age) residual sera following microbiological or biochemical investigations (Osborne et al., 2000), and tested as part of the European Commission-funded second European Sero-epidemiological Network (ESEN2) (Melegaro et al., 2011; Nardone et al., 2007). Children under 5 years old were oversampled with the sample size in each age group ranging from 117 to 192. For those aged 5–20 years approximately 100 sera in each one-year age group were tested. All serological tests for VZV-specific IgG were performed at Preston Public Health Laboratory using a commercial ELISA assay.

Population data were obtained from national statistics offices in the five countries considered as in previous analyses (Melegaro et al., 2011).

2.2. Cluster analysis of social contact data

A cluster is defined as a group of social contacts whose members are more similar to each other than to non-members. The similarity of two contacts is defined using the Manhattan distance measure between the two (Everitt et al., 2011), based on four contact characteristics: proximity (physical, non-physical), duration (<5 min, 5–15 min, 15 min to 1 h, 1–4 h, >4 h), frequency (daily, weekly, monthly, a few times a year, first time) and location (home, work, school, leisure, transport, other). Variables representing each characteristic were recoded so that the distance between any two contacts lies between 0 and 1, so the Manhattan distance becomes the Gower's similarity coefficient which is an appropriate proximity measurement for mixed data (Gower, 1971) (see Appendix A1 for details in Supplementary data). Contacts recorded as taking place in multiple locations were assigned to a single location based on the following hierarchy: home > work > school > leisure > other > transport. The