Mathematisches Forschungsinstitut Oberwolfach

Report No. 55/2013

DOI: 10.4171/OWR/2013/55

Design and Analysis of Infectious Disease Studies

Organised by
Martin Eichner, Tübingen
Elizabeth Halloran, Seattle
Philip O'Neill, Nottingham

10 November – 16 November 2013

Abstract. The fourth workshop on this theme is devoted to the statistical problems of planning and analyzing studies in infectious disease epidemiology.

Mathematics Subject Classification (2000): 92B.

Introduction by the Organisers

Technological advances in recent years have led to the opportunity to routinely collect highly detailed data which can be used to improve our understanding and control of infectious disease spread. This in turn created a need for novel mathematical modelling and statistical analysis. This workshop was based around this broad area, and included two special focus topics, namely molecular typing data and networks. Modern molecular typing techniques have become cheaper and more widespread, with the consequence that field data from outbreaks of infectious disease are increasingly likely to contain highly detailed genome sequence information. The mathematical challenges include integrating complex sequence data into phylogenetic analyses combined with information on disease spread within small clusters and large communities over time. Integration of such methods can produce more detailed insight to transmission chains in populations. A related problem is in understanding the evolutionary selective pressure that the vaccines put on pathogens. Recent developments in data gathering have increased the potential of graph models and network theory to understand which aspects of contact structure are essential for the spread of epidemics and for the planning of control measures. Integrating the theory of network structures with sequence data to refine understanding of actual transmission patterns stands at the forefront of

research on transmission dynamics of infectious diseases. This workshop has focused on the mathematic, statistical, and algorithmic challenges posed by the new technologies for understanding the spread of infectious diseases and for planning and analyzing studies of interventions.

Workshop: Design and Analysis of Infectious Disease Studies

Table of Contents

Abstracts

Emerging epidemics among a population of households

Frank Ball

(joint work with Lorenzo Pellis, Laurence Shaw, and Pieter Trapman)

We consider a stochastic SIR (susceptible → infective → removed) model for the spread of an epidemic among a population partitioned into households. In the first part of the talk, which is based on joint work with Lorenzo Pellis and Pieter Trapman, we use branching processes to define the basic reproduction number R_0 for this epidemic model (Pellis *et al.* [1]) and show inequalities comparing R_0 with previous reproduction numbers for this model. The comparisons imply that, unless all households have size ≤ 3, vaccinating a fraction $1 - R_0^{-1}$ of the population with a perfect vaccine is insufficient to be sure of preventing a large outbreak and they lead to sharper, easily computed bounds for the critical vaccination coverage than were previously available.

In the second part of the talk, which is based on joint work with Laurence Shaw, we consider estimation of the within-household infection rate λ_L from data on the number of removed cases in households during the emerging, exponentially growing phase of an epidemic. We show that using the usual final size distribution for an epidemic in a single household leads to λ_L being appreciably underestimated and use multitype branching process theory to develop an asymptotically unbiased estimator.

References

[1] L. Pellis, F. Ball and P. Trapman, *Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R_0*, Mathematical Biosciences **235** (2012), 85–97.

Patterns of cross-reacting antibodies against influenza A viruses revealed by protein micro array

Michiel van Boven

(joint work with Dennis te Beest, Erwin de Bruin, Sandra Imholz, Jacco Wallinga, Peter Teunis, Marion Koopmans)

Epidemics of influenza A differ substantially in size and in the age distribution of cases. These differences result from varying levels of pre-existing immunity for the strain that is seeding the epidemic or pandemic. In this presentation I will estimate patterns of infection and immunity before and after the pandemic of 2009 by analysis of two cross-sectional population-based serological surveys. Samples were analyzed by a protein microarray test using hemagglutinin (HA1) of different influenza A viruses as antigen. It appears that the microarray is more sensitive than the hemagglutination inhibition assay, is able to provide estimates both of

infection rates and levels of immunity, and allows for a consistent classification of individual sera. Specifically, mixture model analysis of the micro array data yield high estimates of the attack rate in children under the age of ten years and low attack rates in older adults. Bivariate analyses of micro array data yield comparable attack rate estimates, and enable consistent and precise classification of individual samples as belonging to immune or recently infected persons. We show that a high ratio of A/2009 to A/1918 antibody concentrations is indicative of recent infection, while a low ratio is indicative of pre-existing immunity, even if the A/2009 response is high. Population estimates of immunity are low in children aged 5-9 years, and increase monotonically with age. I will discuss the implications for the role of antibody mediated immunity shaping influenza A epidemic patterns.

Inference in epidemics with different types of data, with a view towards genetics

Tom Britton

Classical data for making inference of epidemics relies on observing diagnosis of all or a sample in some community together with information of community structure like households and similar. Quite often there is also temporal information such as date of symptoms and similar. Recently, sequencing infected individuals, using phylogenetic analysis, enables estimation of the transmission tree. In the current ongoing project we investigate how much is gained by combining these types of data. The conclusion seems to be that, also observing the transmission tree enables inference for new parameters and possible deviations from a model, but that not so much is gained if the applied model is correct. Methodology involves likelihood and martingale theory.

Integrating phylogenetics and epidemiology in the study of HIV

Simon D.W. Frost

Human immunodeficiency virus type 1 (HIV-1) is perhaps the most widely studied organism in viral phylodynamic studies, which aim to combine the epidemiology of viral transmission with the evolution of the virus, and for good reason. Not only is HIV-1 infection a significant public health issue, but a vast amount of sequence data has been generated over a period of decades, which when combined with the clock-like nature of HIV-1 evolution, allows fairly accurate reconstruction of past transmission events. Insights that have been gained by HIV-1 sequence analysis include identifying the timing of origin of HIV-1 epidemics, identifying period of exponential epidemic growth, and identification of transmission clusters of infection. To date, rather simple models underlie the analysis of viral sequence data, which consider neither the detailed natural history of viral infection, nor the biased way in which sequence data is typically collected. Moreover, while standard epidemiological models consider dynamics going forwards in time, standard coalescent models for evolution consider dynamics backwards in time. Using HIV

as an example of a model phylodynamic system, I will show how the shape and structure of the phylogenetic tree can be modelled using simple compartmental epidemiological models. These present insights into the meaning of the population genetic effective population size or N_e for epidemics. I show how the peak of N_e may not coincide with the peak of infected individuals, and how the level of asymmetry in the phylogenetic tree may be influenced by heterogeneity in infectiousness at different stages of infection, and by the presence of a core group. I will also present stochastic versions of these models that try to reconcile the forwards-time nature of epidemiological models and the backwards-time nature of coalescent models, and will show how the likelihood of a series of coalescent intervals under a stochastic model may be estimated via constrained simulations. Finally, I will focus on some of the many challenges in phylodynamics, including how to incorporate recombination, selection, and within-host evolution.

Untangling human and animal transmission cycles of vector-borne infections

Sebastian Funk

Many vector-borne infections can be transmitted between animals and humans. The epidemiological roles of different species can vary from important reservoirs to dead-end hosts. Here, we present a method to identify transmission cycles in different combinations of species from epidemiological and ecological field data.

We applied this method to synthesise data from Bipindi, Cameroon, a historical focus of gambiense Human African Trypanosomiasis (HAT, sleeping sickness), a disease that has often been considered to be maintained mainly by humans. We estimated the basic reproduction number R0 of gambiense HAT in Bipindi and evaluated the potential for transmission in the absence of human cases. We found that under the assumption of random mixing between vectors and hosts, gambiense HAT could not be maintained in this focus without the contribution of animals. When using the distributions of species among habitats to estimate the amount of mixing between those species, we found indications for an independent transmission cycle in wild animals. This suggests that elimination strategies may have to be reconsidered as targeting human cases alone would be insufficient for control, and reintroduction from animal reservoirs would remain a threat.

Our approach is broadly applicable and could reveal animal reservoirs critical to the control of other infectious diseases.

On the design of data surveillance strategies for control of epidemics

Gavin Gibson

This talk describes joint work with Max Lau, George Streftaris and Hola Adrakay (Heriot-Watt University) and Glenn Marion (BioSS) on the use of latent residuals in the Bayesian framework to test and compare spatio-temporal models for epidemics and to investigate the effectiveness of momitoring remival strategies. A

key issue in modelling epidemics in space and time is to be able to distinguish the spatial transmission kernel, which characterises the spatial nature of the dispersal of the pathogen from infective to susceptible individuals. Knowledge of the spatial transmission kernel is important when selecting potential control strategies, for example based on eradication of susceptibles within a radius of a newly discovered infection. It is challenging to compare the fit of models that employ different spatial transmission kernels and techniques such as Bayesian model choice can be complicated to implement. In this talk we present an approach that extends the concept of model testing using posterior predictive model checks. Specifically we show how it is possible, using non-centred parameterisations, to define latent processes to which classical tests can be applied in order to assess the validity of the selected model. We show that the process can be defined so that the resulting tests are sensitive to misspecification of the spatial kernel. Moreover we show how the imputed process can be used to provide diagnostics of the nature of the mismatch between model and data. The methods are illustrated using simulated data sets and using a data set on the spatio-temporal dynamics of an invasive plant species in the UK. We also consider how non-centred parameterisations can be used to reduce the amount of simulation required to compare competing control strategies. By comparing the effectiveness of different control strategies applied to realisations of epidemics that share the same system parameters and latent residuals, we describe work in progress that suggests that such a coupling approach has the potential to deliver more efficient estimation of the expected difference in outcomes between control strategies.

Missing dimension in measures of disease intervention impacts

M. Gabriela M. Gomes

(joint work with Marc Lipsitch, Gael Kurath, Andrew R. Wargo, Graham F. Medley, Carlota Rebelo, Antonio Coutinho)

Immunological protection, acquired from either natural infection or vaccine, varies between hosts, reflecting underlying biological mechanisms and affecting population level protection. Distributions of susceptibility and protection entangle with pathogen dose in a way that can be decoupled by adequately representing the dose dimension in the study design. The two extreme distributions of vaccine protection have been termed leaky (equally protects all hosts) and all-or-nothing (totally protects a proportion of hosts). Leaky vaccines are predicted to allow greater pathogen prevalence [1]. Leaky protection induced by natural pathogen exposure generates a threshold in transmission above which infection persists almost unchanged in a population where everyone has been immunized, contrasting with the all-or-nothing regime where population effects of immunity are more noticeable and less sensitive to baseline transmission [2]. These extreme distributions can be distinguished in vaccine field trials from the time-dependence of infections [3]. Frailty mixing models have been proposed to estimate the distribution of protection from time to event data [4]. However, results are not comparable across

regions unless there is explicit control for baseline transmission [5]. We provide a rationale for how comparability can be attained, and trial efficiency enhanced, by adopting study designs and estimation procedures that integrate multiple populations, covering a wide range of transmission intensities. Distributions of host susceptibility, and acquired protection, can be estimated from dose-response data generated under controlled experimental conditions [6]–[7] and natural settings [8]. These distributions can guide research on mechanisms of protection, as well as enable model validity across the entire range of transmission intensities. We argue that a shift to a dose-dimension paradigm is urgently needed in infectious disease science and public health.

References

[1] Goldstein E, Paur K, Fraser C, Kenah E, Wallinga J, Lipsitch M *Reproductive numbers, epidemic spread and control in a community of households*, Math Biosci **221** (2009), 11–25.

[2] Gomes MGM, White LJ, Medley GF *Infection, reinfection, and vaccination under suboptimal immune protection*, Epidemiological perspectives. J Theor Biol **228** (2004), 539–549.

[3] Smith PG, Rodrigues LC, Fine PE *Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies*, Int J Epidemiol **13** (1984), 87–93.

[4] Halloran ME, Longini Jr IM, Struchiner CJ *Estimability and interpretability of vaccine efficacy using frailty mixing models*, Am J Epidemiol **144** (1996), 83–97.

[5] Struchiner CJ, Halloran ME *Randomization and baseline transmission in vaccine field trials*, Epidemiol Infect **135** (2007), 181–194.

[6] Furumoto WA, Mickey R *A mathematical model for the infectivity-dilution curve of tobacco virus*, Theoretical considerations. Virology **32** (1967), 216–223.

[7] Haas CN, Rose JB, Gerba CP *Quantitative Microbial Risk Assessment*, John Wiley & Sons, Inc, New York (1999).

[8] Smith DL, Dushoff J, Snow RW, Hay SI *The entomological inoculation rate and Plasmodium falciparum infection in African children*, Nature **438** (2005), 492–495.

Sequential Monte Carlo methods for inferring transmission dynamics from pathogen genetic sequences

Edward Ionides

Sequential Monte Carlo (SMC) algorithms [6]–[2] typically operate on partially observed Markov process (POMP) models. The Markov process $\{X(t)\}$ may take values in an arbitrary space χ, with transition probabilities depending on a real vector-valued parameter Θ. Data y_1, ...,y_N are obvserved at times t_1,...,t_N and are modeled as being realizations of random variables Y_1,...,Y_N which are conditionally independent given $\{X(t)\}$ having conditional density $f_{Y_n|X_n}(\cdot|\cdot;\Theta)$. POMP models have repeatedly been proposed as a general framework for modeling biological systems,since they provide a reasonable tradeoff between generality and tractability. In the context of phylodynamic inference, the state of the Markov process may be a tree, with branches and leaves being added through time. SMC methodology was proposed independently by multiple groups in the 1990s, simultaneously called particle filtering, bootstrap filtering, Monte Carlo filtering, sequential importance sampling, and the condensation algorithm. It was used in physics and

chemistry since the 1950s: marketed as "Poor man's Monte Carlo" [7]. In modern theory, SMC and Markov chain Monte Carlo (MCMC) have similar asymptotic guarantees. SMC provides an alternative to MCMC for many computations, and for dynamic systems SMC is often preferred. An SMC algorithm consists of a *swarm of particles* evolving according to the stochastic dynamic system. Particles consistent with data are propagated; those inconsistent with data are pruned. The propagation and pruning are done in such a way that SMC approximates an ideal nonlinear filter, and in particular SMC gives unbiased likelihood estimates, with more particles giving reduced variance. A recursive description of a basic SMC algorithm is as follows:

[**Filter at time** n+1]. Suppose inductively that we have a swarm of J particles, $\{X^F_{n,j}, j = 1, ..., J\}$ whose distribution is a good approximation to the conditional distribution of $X(t_n)$ given data up to time t_n. In that case, we say that $\{X^F_{n,j}, j = 1, ..., J\}$ gives a numerical solution to the *filtering* problem at time t_n.

[**Predict at time** n + 1]. Each particle $X^F_{n,j}$ is simulated forward from time t_n to time t_{n+1}. The resulting swarm, denoted $\{X^P_{n+1,j}, j = 1, ..., J\}$ represents the distribution of $X(t_{n+1})$ given data up to time t_n. We say that $\{X^P_{n+1,j}, j = 1, ..., J\}$ is a numeric solution to the *prediction* or *1-step prediction* or *1-step forecasting* problem at time t_{n+1}.

[**Filter at time** n+1]. Drawing a random sample of size J with replacement from $\{X^P_{n+1,j}, j = 1, ..., J\}$, with the sampling weight for $X^P_{n+1,j}$ proportional to $f_{Y_{n+1}|X_{n+1}}(y_{n+1}|X^P_{n+1,j}; \Theta)$ gives a swarm $\{X^F_{n+1,j}, j = 1, ..., J\}$ which is a numerical solution to the filtering problem at time t_{n+1}.

SMC gets the statistician only part of the way toward inference on unknown parameters or model selection: it estimates latent dynamic variables, and integrates them out to approximate the likelihood, for a given model and given parameter values, and using these noisy and computationally expensive likelihood evaluations in Bayesian or frequentist inference is tricky. The issue has inspired much research over the past decade. Iterated filtering [10],[9] aims to maximize the likelihood by adding a random walk in parameter space. Particle MCMC [1] aims to simulate a posterior distribution by using an SMC estimate of the likelihood in an MCMC investigation of the parameter space. Iterated filtering and particle MCMC both inherit from SMC an important property: they require simualation from the dynamic model but they do not require explicit computation of transition probabilities. This is called the *plug-and-play* property [5],[8]. Plug-and-play facilitates model development (we can simualate from many models of interest for which transition probabilities are hard to compute). Plug-and-play also facilitates software development: inference software, such as the R package **pomp** [11], can simply takes as an argument code to generate simulations. Other plug-and-play methods, such as synthetic likelihood [18] and approximate Bayesian computation (ABC) [17] have been developed recently. Iterated filtering, building on SMC, allows effective inference for a wide range of disease transmission models: examples include malaria [13], measles [8] and cholera [12]. References to many more

examples are available at *http://en.wikipedia.org/wiki/Iterated_filtering*. Spatial-temporal and high dimensional systems remain a challenge for SMC. Genetic data is a new frontier. Rasmussen et al. [15] developed a method for SMC-based inference conditional on a phylogeny, by showing that if uncertainty in the phylogeny is negligible then a coalescent process on this phylogeny can be used as a measurement model for applying SMC techniques to stochastic dynamic transmission models. This reduces the problem to nonlinear time series analysis, where the data are a time series of the number of coalescent times in small, discrete time intervals. Some uncertainty in the estimated phylogeny can be accounted for, but mutually consistent estimation of the phylogeny and the transmission model is currently unresolved. Bouchard-Côté et al. [4] used SMC with a tree-valued Markov process to estimate a phylogeny, by building a phylogeny backwards in time, so a "particle" is a forest of trees that combine down to a single large tree as the filtering proceeds. SMC for joint estimation of transmission dynamics and phylogeny is also possible, as shown by preliminary results (in collaboration with Alex Smith and Aaron King). Each particle is a transmission tree of all infected individuals in a population. Tree growth follows the forward-time dynamic model for disease tranmission. Observations are assignments of sequences to branches on the tree. Currently, we can filter simulated data from simple models with, say, 100 observed sequences. Some improvements are expected through refining the code, but an important open question is whether fundamental algorithmic developments could enable, say, 1000 sequences and 20-parameter models. This will involve grappling with the curse of dimensionality, which in the context of SMC is the problematic fact that the basic SMC algorithm becomes exponentially more challenging numerically as the dimension of the state increases [3]. Hope for resolving the curse can be found in the observation that SMC could in principle also require a number of particles exponential in the length of the time series. This is infeasible, but is avoided when the Markov process has *temporal mixing* properties [14]. Recent work by Rebeschini and van Handel [16] shows the possibility of developing new SMC algorithms that take advantage of weak spatial coupling to avoid the curse. In the phylodynamic context, weak coupling arises when lineages interact only through competition for susceptibles, and we are investigating the construction of practical algorithms based on the advances of [16]. It has been said that genetic data 'decorrelate' the lineages, but ecological competition still exists and can be critical.

References

[1] Andrieu, C., Doucet, A., and Holenstein, R. *Particle Markov chain Monte Carlo methods*, Journal of the Royal Statistical Society, Series B (Statistical Methodology), **72** (2010), 269–342.

[2] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. *A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking*, IEEE Transactions on Signal Processing, **50** (2002), 174–188.

[3] Bengtsson, T., Bickel, P., and Li, B. *Curse-of-dimensionality revisited: Collapse of the particle Filter in very large scale systems*, In Speed, T. and Nolan, D., editors, Probability

and Statistics: Essays in Honor of David A. Freedman, Institute of Mathematical Statistics, Beachwood, OH (2008), 316–334.
[4] Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. *Phylogenetic inference via sequential Monte Carlo*, Systematic Biology **61(4)** (2012), 579–593.
[5] Bretó, C., He, D., Ionides, E. L., and King, A. A. *Time series analysis via mechanistic models*, Annals of Applied Statistics **3** (2009), 319–348.
[6] Doucet, A., de Freitas, N., and Gordon, N. J., editors *Sequential Monte Carlo Methods in Practice*, Springer, New York.
[7] Hammersley, J. M. and Morton, K. W. *Poor man's Monte Carlo*, Journal of the Royal Statistical Society, Series B (Statistical Methodology) **16** (1954), 23–38.
[8] He, D., Ionides, E. L., and King, A. A. *Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study*,Journal of the Royal Society Interface **7** (2010), 271–283.
[9] Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. A. *Iterated filtering*, Annals of Statistics **39** (2011), 1776–1802.
[10] Ionides, E. L., Bretó, C., and King, A. A. *Inference for nonlinear dynamical systems*, Proceedings of the National Academy of Sciences of the USA **103** (2006), 18438–18443.
[11] King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S., and Kendall, B. *pomp: Statistical inference for partially observed Markov processes* (2009).
[12] King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. *Inapparent infections and cholera dynamics*, Nature **454** (2008), 877–880.
[13] Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Yadav, R., Dhiman, R., and Pascual, M. *Forcing versus feedback: Epidemic malaria and monsoon rains in NW India*, PLoS Computational Biology **6** (2010), e1000898.
[14] Del Moral, P. and Guionnet, A. *On the stability of interacting processes with applications to filtering and genetic algorithms*, Annales de l'Institut Henri Poincare (B) Probability and Statistics**37(2)** (2001), 155–194.
[15] Rasmussen, D. A., Ratmann, O., and Koelle, K. *Inference for nonlinear epidemiological models using genealogies and time series*, PLoS Computational Biology **7(8)** (2011), e1002136.
[16] Rebeschini, P. and van Handel, R. *Can local particle filters beat the curse of dimensionality?*, Arxiv **1301.6585** (2013).
[17] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems*, Journal of the Royal Society Interface **6** (2009), 187–202.
[18] Wood, S. N. *Statistical inference for noisy nonlinear ecological dynamic systems*, Nature **466** (2010), 1102–1104.

Survival analysis, who-infected-whom, and phylogenetics in infectious disease epidemiology

Eben Kenah

Statistical methods for infectious disease data based on generation or serial intervals fail to account for information contributed by uninfected person-time and for competing risks of infection. These problems are solved by methods based on contact intervals, which are times between the onset of infectiousness and infectious contact. These allow parametric, nonparametric, and semiparametric analyses of infectious disease transmission data using methods adapted from survival analysis. When who-infects-whom is not observed, these estimators are sums or averages

over all possible transmission trees. A phylogenetic tree linking pathogen genetic sequences provides partial information about who-infected-whom. Here, we describe an algorithm for finding all possible transmission trees consistent with epidemiologic data and a pathogen phylogeny.

Bayesian non-parametric inference for epidemic models

Theodore Kypraios

Despite the enormous attention given to the development of methods for efficient parameter estimation, there has been relatively little activity in the area of non-parametric inference. That is, drawing inference for the quantities which govern transmission, i) the force of infection and ii) the period during which an individual remains infectious, without making certain modelling assumptions about its (parametric) functional form or that it belongs to a certain family of parametric distributions. In this talk we describe three approaches which allow Bayesian non-parametric inference for the force of infection; namely via Gaussian Process, Step Functions, and B-splines. We illustrate the proposed methodology both with simulated and real datasets.

The rise and fall of HIV in Africa: A challenge for mathematical modeling

Nico J.D. Nagelkerke

Many parts of sub-Saharan Africa have experienced large HIV epidemics with rapid onsets, generally attributed to a combination of factors related to high risk sexual behavior. In several countries the adult prevalence of HV exceeded 20%. In the rest of the world, even places where the same (sexual) risk factors appeared to be present, the HIV prevalence in the general population has remained substantially lower. In several African countries, however, HIV prevalence and incidence are declining rapidly; declines that began prior to widespread therapy or implementation of any other major biomedical prevention. This change has been construed as evidence of behavior change, but direct evidence for behavior changes of an extent and timing that would explain this decline is lacking. Here, we look at the structure of current mathematical models and argue that the common "fixed risk per sexual contact" assumption implies the conclusion of substantial behavior changes. We argue that this assumption ignores reported non-linearities between exposure and risk. Taking this into account we propose that some of the decline in HIV transmission may be part of the natural dynamics of the epidemic, and that several factors that have traditionally been ignored by modelers for lack of precise quantitative estimates may well hold the key to understanding epidemiological trends. Most importantly heterogeneity in susceptibility has been largely ignored, despite empirical evidence for its importance. This heterogeneity not only changes our understanding of the course of the epidemic but also has a bearing on the design of transmission studies.

MCMC for a birth-death-mutation (BDM) model

Peter Neal

A birth-death-mutation (BDM) model has been used by a number of authors to model the evolution of a tuberculosis epidemic in San Francisco in the early 1990s. The observed data is assumed to be a cross-sectional study of the tuberculosis outbreak. It is impossible to write down the likelihood for the model without substantial, non-trivial data augmentation which prohibits the use of standard MCMC algorithms. However it is trivial to simulate a realisation of the BDM model and ABC algorithms have been used to estimate the parameters of the BDM model. Starting from the ABC perspective that simulation is straightforward, we construct an MCMC algorithm which uses simulation. Specifically we use a non-centered parameterisation which enables us to treat the simulation process as a data augmentation problem and takes similar amounts of time per iteration as the ABC algorithms. The MCMC algorithm is successfully applied to the San Francisco tuberculosis data.

How much epidemiology can we infer from phylogeny

David A. Rasmussen

Population geneticists have long recognized that genealogies contain useful information about the demographic history of a population, especially with respect to past population dynamics and population structure. This has lead to the development of several coalescent-based methods for inferring historical demography from genealogies. The coalescent in essence provides a probability distribution on trees under a certain demographic model, and therefore allows likelihood or Bayesian inference of demographic parameters from genealogies. Coalescent methods can also be applied to genealogies of infectious pathogens, allowing epidemiological dynamics and parameters to be inferred from sequence data. While it is well understood how different demographic processes like population growth or population structure affect the shape of genealogies, it is much more difficult to quantify how much information genealogies contain about demography. I propose that genealogies are in general highly informative about past population dynamics, and therefore it should be possible to infer past demographic changes relatively easily from genealogies. On the other hand, estimating parameters relating to population structure precisely from genealogies may be much more difficult, especially in weakly structure populations. As an illustrative example, I use the case of dengue serotype 1 (DENV-1) to show how much information can be obtained from genealogies about historical epidemiological dynamics. Dengue provides an interesting case study because earlier studies found that population dynamics estimated from dengue genealogies correspond poorly with dengue dynamics observed in hospitalization data. Earlier work had failed to detect any signal of dengues strong seasonal dynamics in southern Vietnam, suggesting that there is insufficient information in genealogies to estimate population dynamics on the relevant time scale. I show that the inability to accurately estimate dengues seasonal dynamics may actually

be due to using a coalescent model for inference that does not properly take into account the spatial structure of the dengue population. Using coalescent models that take into account spatial structure, I show that it is possible to reconstruct historical dynamics of DENV-1 consistent with patterns in time series data. Estimating parameters relating to population structure from genealogies appears to be more difficult than estimating historical dynamics. I demonstrate this using genealogies simulated under different demographic scenarios, varying the strength of seasonality by varying how much mixing occurs between different subpopulations in the demographic model. These simulations show that if population structure is strong, the structure of the population strongly impacts the shape of the genealogy and so it is relatively easy to estimate parameters relating to population structure from the genealogies. On the other hand, if population is weak, lineages in the genealogy will move rapidly between populations. The rapid movement of lineages between populations means that observing the location or state of a lineage at the time of sampling provides little information about the past state of the lineage. There is therefore a high degree of uncertainty in the state of a lineage over the majority of the tree, which makes population structure difficult to estimate from a genealogy. However, how much information a genealogy contains about population structure can be highly dependent on the sample size. Increasing the sample size can greatly improve our ability to estimate parameters relating to population structure. Therefore, while it may be possible to reconstruct population dynamics from a limited number of sampled lineages, inferring the structure of a population from genealogies may require much larger sample sizes.

How epidemiology interacts with ecology

Mick Roberts
(joint work with Hans Heesterbeek)

It has been a paradigm in ecology that large complex ecosystems tend to be unstable [3], although the nature of the interaction between species is a strong determinant of persistence [1, 4]. Recently, Dobson speculated that "Parasites look increasingly viable as the 'missing links' in food webs, the 'dark matter' that helps stabilize otherwise unstable structures (presentation at Isaac Newton Institute for Mathematical Sciences, 20/8/2013).

We have developed a differential equation model that describes how infectious diseases may alter the interactions between populations in a complex food web [5]. The stability matrix at the infection-free steady state partitions into a community matrix, that determines ecological stability, and a matrix that determines epidemiological stability. Furthermore, this second matrix may be used to construct the next generation matrix in the usual way [2], and to determine which species are reservoirs of infection.

Two simple examples illustrate this approach. The first of these is a model of the rinderpest, wildebeest, grass interaction, where the inferred dynamics qualitatively match the observed phenomena that occurred after the eradication of rinderpest

from the Serengeti ecosystem in the 1980s. The second example is a prey-predator system, where both species are hosts of the same pathogen. It is shown that regions for the parameter values exist where the two host species are only able to coexist when the pathogen is present to mediate the ecological interaction.

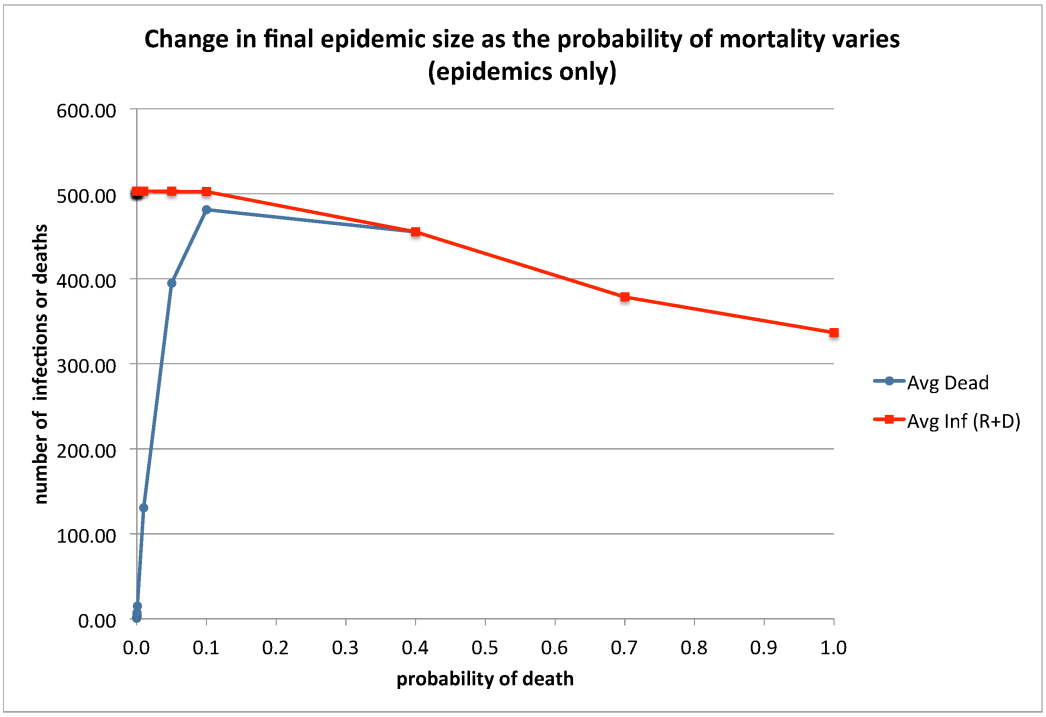
References

[1] Allesina S, Tang S. 2012 Stability criteria for complex ecosystems. Nature 483:205-8.
[2] Diekmann O, Heesterbeek JAP, Roberts MG. 2010 The construction of next-generation matrices for compartmental epidemic models. Journal of the Royal Society Interface 7:873-85.
[3] May RM. 1972 Will a large complex system be stable? Nature 238:413-4.
[4] Neutel A-M, Heesterbeek JAP, de Ruiter PC. 2002 Stability in real food webs: weak links in long loops. Science 296:1120-3.
[5] Roberts MG, Heesterbeek JAP. 2013 Characterizing the next-generation matrix and basic reproduction number in ecological epidemiology. Journal of Mathematical Biology 66:1045-64.

The impact of disease-related mortality during epidemics in small communities

Lisa Sattenspiel

When populations are very small, it is difficult to design and analyze appropriate mathematical models to address important questions about the spread of disease at the population level. One relatively new approach to this problem is to develop agent-based computer simulation models that can incorporate the most essential characteristics of the population as well as the stochasticity that is so important for small populations. This talk described the structure of an agent-based model designed to study the spread of the 1918-19 influenza pandemic in a small fishing community in the Canadian province of Newfoundland and Labrador. The talk also presented results from sensitivity analyses of the model, with a focus on the impact of disease-related mortality. Sensitivity results discussed in the talk included a comparison of different versions of the model, including versions with a) completely random movement of individuals on the model space, b) partial directed movement where fishermen and children moved on a daily basis to boats and school, respectively, while all other agents moved within their houses only, c) full directed movement, where all agents moved among different locations within the simulated community according to a specified timetable and movement scheme, and d) full directed movement with disease-related mortality added to the model. Results showed that both full and partial directed movement resulted in much faster epidemic spread and higher levels of infection than did random movement. One initially surprising result is that simulations using mortality levels characteristic of the 1918-19 influenza pandemic on the island of Newfoundland differed very little from those using the same type of movement but no mortality. This is due to the fact that the modeled community is so small (503 people) that only 4 deaths would be expected at the observed rate of mortality. An analysis of the impact of higher probabilities of death indicates that at very low levels of mortality



and with high probabilities of transmission, nearly everyone becomes infected, but very few die. The proportion of cases that die increases rapidly as mortality levels increase, however, and at high levels of mortality, nearly everyone infected dies. However, because transmission chains end with the death of an agent, in this case a significant fraction of agents never become infected (see figure). These results suggest that the major impact of a disease like influenza is less a consequence of disease-related mortality than it is of the extent of disease-related illness and the disruptions in life due to the large number of cases that can occur and the alterations in behavior that accompany illness.

Quantifying the relationship between shedding and transmission for *Escherichia coli* O157 in feedlot cattle

Simon E. F. Spencer, Thomas E. Besser, Rowland Cobbold, Nigel P. French

Abstract. In this study we were able to quantify the relationship between shedding and transmission of *E. coli* O157 in cattle. We achieved this by fitting a non-Markovian stochastic SIS (Susceptible-Infected-Susceptible) transmission model that accounts for indirect infections from historic shedding. By considering information from multiple tests we were able to simultaneously estimate the test sensitivities alongside the unobserved colonisation status for each individual animal throughout the study. Our results were based on a 99 day longitudinal study of 160 animals undergoing natural infection and housed within a commercial feedlot setting.

In the literature much has been made of the role that 'supershedding' individuals might play in the transmission dynamics of *E. coli* O157. However we found little evidence of a disproportionate contribution to transmission from high-shedding individuals. We were able to learn further about the transmission dynamics from a subset of *E. coli* samples for which additional genetic typing data were available. There was clear evidence of strains being transmitted within pens as well as the invasion of new strains from outside the pens. Since all of the samples remain frozen in storage, we have been able to design and carry out an additional study that utilizes this kind of genetic information more effectively.

Introduction. Shiga toxin-producing Escherichia coli O157 (STEC O157) continue to present a serious threat to public health in many countries around the world[1]. The severe haemorrhagic diarrhoea and serious sequelae associated with infection can result in loss of life[2] and, although STEC O157 are associated with a relatively low incidence compared to other enteric zoonoses, they present a major burden to the economies of many countries.

There has been much discussion in the literature of the role and importance of so called supershedders in transmission and the maintenance of infection in cattle[1]. These are defined as animals shedding relatively large concentrations of STEC O157 in their faeces, usually between 10^3 and 10^4 /g for an extended period[3]. They are considered to be responsible for a disproportionate amount of transmission.

In this study, we present the first formal consideration of the relationship between the concentration of STEC O157 shed in faeces and transmission, using the results from a large-scale study of natural infection in feedlot cattle. We fitted SIS (Susceptible-Infected-Susceptible) models that consider in detail the relationship between shedding and transmission, as well as the diagnostic test characteristics. We derive the shape of the relationship between shedding and transmission and provide new insight into the role played by supershedders in the maintenance of infection in cattle populations.

Methods. The colonisation status of each animal was measured twice weekly using two diagnostic tests: a Recto-Anal Mucosal Swab (RAMS) and a faecal pat sample. Each sample was tested for *E. coli* O157 using culture and PCR and the concentration of bacteria was estimated. Although we assumed perfect test specificity, we allowed for the fact that both tests had less than perfect sensitivity.

Markov chain Monte Carlo techniques were used to fit a discrete-time SIS transmission model to these data. This involved inferring the hidden colonisation status of each animal on each day of the study and enabled the diagnostic test sensitivities to be estimated. By incorporating the shedding levels into the risk of colonisation we were able to estimate directly the shape of the relationship between them.

In particular we assumed that the rate of colonisation had the form $\alpha + \beta \sum_i S_i^\eta$, where S_i is the level of shedding from individual i in the pen, measured in log10

cfu. The parameters α and β balance the risk of colonisation from outside the pen and from individuals within the pen. The parameter η governs the nature of the relationship between shedding and transmission. If the risk of transmission was proportional to the number of bacteria shed we would expect $\eta >> 1$, whilst if colonisation status rather than shedding level was the risk factor for transmission then we would expect $\eta = 0$.

We adapted this model further to account for the risk of infection from historical shedding. We assumed that a fixed proportion of bacteria survive each day in the environment and these bacteria also pose a risk of colonisation. Notice that these assumptions yield a non-Markovian model that allows infectious material to accumulate in the environment which can lead to colonisations later in the study.

Results. The posterior distribution of η was largely below one, providing little evidence that high shedding individuals contribute disproportionately to the burden of infection. We also inferred a surprisingly short duration of survival of the bacteria within the environment. Overall we found that the animals must shed at levels over 10^4 cfu/g before it is possible for the infection to persist in the pen without reintroductions from outside. Since the pens used in the study were of two different sizes we were able to infer that animal density had a strong impact on transmissibility, which could have future implications for control.

Genetic typing data produced via Pulsed Field Gel Electophoresis (PFGE) showed evidence of a single strain of *E. coli* O157 circulating within a pen as well as occasions where a new strain invaded from outside. The original subset of samples chosen for genotyping were not selected at random, potentially introducing bias into the analysis. Since the sampled material remains in frozen storage, we have been able to design and undertake a more comprehensive genotyping study, the results from which should be available shortly.

Discussion. The main aim of this study was to quantify the relationship between bacterial shedding levels and the risk of colonisation. We have not been able to substantiate the hypothesised link between supershedders and high transmission. In future we aim to use model comparison techniques to formally assess this hypothesis.

The original supershedding hypothesis was based on the idea that persistently high shedding individuals were colonised by the bacteria at the recto-anal junction (RAJ), whilst low shedding individuals resulted from amplification of the bacteria during passage through the digestive tract. Supershedders would therefore constitute an identifiable subgroup of individuals at which control measures could be targeted. However, the lack of a relationship between high shedding and excess transmission, coupled with the large number of animals that tested positive in the RAJ swabs despite low bacterial levels in faeces, suggests that this hypothesis now needs to be reconsidered.

References

[1] Chase-Topping M *et al.* (2008). Supershedding and the link between human infection and livestock carriage of Escherichia coli O157. *Nature Reviews*, **6**, 904–912.
[2] Karmali, MA (2004). Infection by shiga toxin-producing *Escherichia coli*. *Molecular Biotechnology*, **26**, 117–122.
[3] Omisakin F *et al.* (2003). Concentration and prevalence of *Escherichia coli* O157 in cattle feces at slaughter. *Appl. Environ. Microbiol.* **69**, 2444–2447.

Dynamics of influenza and modes of transmission

Nikolaos I. Stilianakis

(joint work with Yannis Drossinos, Marguerite Robinson, and Thomas P. Weber)

The epidemiology of airborne infectious diseases, such as influenza, is characterised by multiple modes of transmission. In the case of influenza three, non-mutually exclusive, modes of transmission have been identified; airborne transmission mediated by respirable droplet nuclei (aerodynamic diameter $d_a \leq 10\mu m$), droplet transmission mediated by inspirable large droplets ($10\mu m \leq d_a \leq 100\mu m$), and contact transmission mediated by droplets settled in the environment. Their relative importance and the efficiency of control measures depend, among other factors, on the inactivation of viruses in different environmental media. On inanimate surfaces and in aerosols daily virus inactivation rates are of the order of 1-100 and they can survive several hours, whereas on hands the inactivation rates are of the order of 1000 and viruses can survive for a few minutes [1]. Understanding the dynamics of transmission of influenza and the relative importance of the associated modes of transmission is of major interest since it would uncover the underlying biological and physical processes, and it can be of use for the design of effective and efficient intervention strategies.

A series of mathematical models that describe the transmission dynamics of influenza in humans explicitly incorporating the modes of transmission, such as transmission via respiratory droplets and contact in space and time, were developed. Droplets dynamics is determined by their physical properties, whereas population dynamics is determined by, among other properties, the pathogen infectivity and the host contact rates. A fundamental time dependent model suggests that airborne infections, mediated by respirable droplets, provides the dominant mode of transmission in middle and long-range epidemics whereas larger, so called inspirable droplets, be they airborne or settled, characterise short-term epidemics with high attack rates [2]. The model has the typical population structure of a susceptible-infected-recovered (SIR) model with the particular feature that it explicitly describes the dynamics of the pathogen-carrying droplets via the above described modes. In its general form the fundamental model can be described as follows.

The emitted polydisperse droplet distribution is divided into l bin, each one characterised by an average droplet diameter d_i, an airborne droplet number $D_i(t)$, and a corresponding number of settled droplets $C_i(t)$. We consider that $d_j \geq d_i$

for $j \geq i$. The dynamics of the infection for such a discretised distribution is described by

$$\frac{dS}{dt} = -\sum_{i=1}^{l} N_p(d_i) \left[\beta_{d_i}\, q_{d_i}(d_i) D_i + \beta_{c_i}\, q_{c_i}(d_i) C_i\right] \frac{S}{N},$$

$$\frac{dI}{dt} = -\frac{dS}{dt} - \mu_I I,$$

$$\begin{aligned} \frac{dD_i}{dt} = \kappa_i I \quad &- \quad \left[\left(\frac{B}{V_{cl}} + \tilde{c}_{d_i}\right) q_{d_i}(d_i) + \mu_d + \theta_i\right] D_i \\ &+ \quad \sum_{j>i} \phi_{ji} D_j - D_i \sum_{j<i} \phi_{ij} \ , i, j = 1, \ldots, l \ , \end{aligned}$$

$$\begin{aligned} \frac{dC_i}{dt} = \theta_i D_i \quad &- \quad \left[\left(\eta + \tilde{c}_{c_i}\right) q_{c_i}(d_i) + \mu_c\right] C_i \\ &+ \quad \sum_{j>i} \tilde{\phi}_{ji} C_j - C_i \sum_{j<i} \tilde{\phi}_{ij} \ , \quad i, j = 1, \ldots, l, \end{aligned}$$

$$\frac{dR}{dt} = \mu_I I,$$

with appropriate initial conditions, e.g., $S(0) = S_0$, $I(0) = I_0$, $D_j(0) = 0$, $C_j(0) = 0$, $R(0) = 0$, and $S + I + R = N$, with N the total, constant, population size and S, I, D, C, and R denoting the populations of susceptible and infected persons, droplet and settled droplets and recovered population, respectively. In the equation for the droplets, D_i, the last two terms are evaporation terms with ϕ_{ji} the evaporation rate of droplet d_j to become droplet d_i; the penultimate term models the increase of D_i droplets due to evaporation of all larger droplets ($j > i$), and the last term its decrease via evaporation to smaller droplets. Similarly, for settled droplets, (equation for C_i), the same evaporation terms for settled droplets are denoted by $\tilde{\phi}_{ij}$. We neglect non-linear processes that convert smaller droplets into larger by coagulation and the inverse process of droplet break up.

Extension of the model to a spatio-temporal mathematical model with droplet transport determined by diffusive and convective processes reveals that respirable droplets can lead to an epidemic wave propagating through a fully susceptible population or a secondary infection outbreak for a localised susceptible population. Droplet diffusion is found to be an inefficient mode of droplet transport leading to minimal spatial spread of infection. A threshold air velocity above which disease transmission is impaired, even when the basic reproduction number R_0 exceeds unity, is derived [3].

In a subsequent model, a modelling approach describing the transmission of airborne infections, such as influenza, that captures the effect of seasonal pathogen inactivation on infection disease periodicity reveals that the introduction of seasonally forced pathogen inactivation rate leads to a time delay between peak pathogen survival and peak diseases incidence. The observed oscillations are found to have a

period identical to that of the seasonally forced inactivation rate, the period being independent of the duration of infection acquired immunity [4].

References

[1] Th. P. Weber, & N.I. Stilianakis, Inactivation of influenza A viruses in the environment and modes of transmission: A critical review, J. Infect. **57**, 361-373 (2008).
[2] N. I. Stilianakis & Y. Drossinos, Dynamics of infectious disease transmission by inhalable respiratory droplets, J.R. Soc. Interface, **7**, 1355 - 1366 (2010).
[3] M. Robinson, N. I. Stilianakis, & Y. Drossinos, Spatial dynamics of airborne infectious diseases, J. Theor. Biol., **297**, 116-126 (2012).
[4] M. Robinson, Y. Drossinos, & N.I. Stilianakis, Indirect transmission and the effect of seasonal pathogen inactivation on infectious disease periodicity, Epidemics **5**, 111-121 (2013).

Population structure, why bother?

Pieter Trapman
(joint work with Frank Ball, Tom Britton, Jean-Stephane Dhersin, Viet Chi Tran, Jacco Wallinga)

In epidemiology we are often interested in R_0 for emerging infectious diseases. We discuss how this quantity depends on the population structure and relatively easy to obtain parameters, such as the shape of the infectivity profile and the Malthusian parameter. We show that population structures such as random (configuration model) networks, household structures and multi-type populations hardly influence the relationship between the Malthusian parameter and R_0.

Inference of who infected whom using genetic data in the presence of incomplete sampling: Applications to the HIV epidemic in MSM

Erik Volz

We consider the genealogical structure generated by common epidemiological models such that each lineage in a genealogy corresponds to a single infected host and each node corresponds to a transmission between hosts. Call this genealogy the 'transmission genealogy'. Under suitable assumptions about the natural history and immunological dynamics of a pathogen, the phylogeny of a pathogen will correspond to the transmission genealogy. We show that this is a valid approximation if super-infection is rare (a host becomes infected once and only once) and if the effective population size of the pathogen within hosts is very small. Under these conditions, we outline a coalescent mathematical model (a model which describes genealogical structure on a retrospective time axis), which relates the genealogical structure predicted by epidemiological models to phylogenies which may be estimated from pathogen genetic sequence data. This theory enables the estimation of compartmental model parameters from pathogen genetic data. We present several applications of this theory. Firstly, we illustrate under what conditions the effective population size of the pathogen at the epidemic level will correspond to the

the true prevalence of infection. We show that the correspondence between effective size and prevalence is good during the early exponential growth period of an epidemic, but that the correspondence can be very bad if the per-capita incidence rate changes rapidly through time. Secondly, we fit two simple compartmental SIR models to a phylogeny of S. aureaus, yielding an estimate of R0 and the early epidemic growth rate. Thirdly, we fit a complex compartmental model for the HIV epidemic to a phylogeny estimated from 662 subtype B HIV-1 sequences. By fitting this model, we estimate that 45% of transmissions likely occur during the first year of the infectious period. Finally, we discuss applications of the coalescent models to the problem of inferring the source of infection (transmission pairs) in large random samples of infected hosts and corresponding pathogen sequences.

The Modeling of Pathogen-Specific Counts for Hand, Foot and Mouth Disease

Jon Wakefield

(joint work with Cici Chen, Leigh Fisher, Steve Self and Betz Halloran)

In this talk modeling approaches for hand, foot and mouth disease (HFMD) data collected in China over the period 2009-2010 will be discussed. HFMD is caused by an acute contagious viral infection and there have been large-scale outbreaks in Asia during the past 20 years. The disease is mostly in children, with fecal-oral transmission. Enterovirus 71 (EV71) and Coxackie A16 (CoxA16) are the most common viruses associated with HFMD. Cases are most infectious during the first week of acute illness but may continue to shed virus in the stool for weeks. The data are collected via a surveillance system, and are made available by the Chinese Center for Disease Control (CDC). The basic data consist of weekly counts in geographical areas, by age and gender and with an indicator of the severity of disease (mild or severe). A certain proportion of mild and severe cases are selected for lab testing, so that the pathogen responsible for disease is known. Enterovirus 71 (EV71) and Coxackie A16 (CoxA16) are the most common viruses associated with HFMD. Meteorological data, and other area-level covariates are also available. The talk has three parts. In the first part, a spline model is described to inform on the medium to large scale spatio-temporal variability. The spline regression coefficients are assigned a space-time smoothing prior and the model is fitted using the integrated nested Laplace approximation (INLA) so that computation is fast. In the second part of the talk, we consider the joint modeling of the basic count and pathogen data. For these data, the total number of severe and mild cases by pathogen are unobserved, but these may be introduced into an MCMC approach as auxiliary variables. The parameters of primary interest are (in a generic area and time period): p=Pr(case), qE = Pr(EV71 — case), rE = Pr(severe — EV71) and rC = Pr(severe — CoxA16). The auxiliary variable scheme is computationally expensive since the unobserved variables have large support. Hence, as an alternative, we derive estimators for the four probabilities of interest, along with an associated asymptotic variance-covariance matrix. The

sampling distribution of the estimator is then used as a likelihood, with a space-time smoothing prior at the second stage of a hierarchical model. The aim then is to model the (logits of the) probabilities as a function of covariates and time, to understand the interplay (competition) between the pathogens. The next step is to introduce space, so that the space-time dynamics can be modeled. Finally, in the third part of the talk, we consider the design of a vaccine trial for EV71 HFMD. In particular we consider matched pairs of areas with individuals in one control area being given placebo, and individuals within the other area being randomized to either placebo or vaccine, according to some fixed fraction. This is an example of a two-stage randomization since areas and individuals are randomized. Within-area dependence in the binary indicator of disease may be dealt with via random effects or sandwich estimation. With this design direct, indirect, total and overall vaccine effects may be estimated with simulations being used to assess the power as a function of different numbers of pairs, and the fractions vaccinated within areas. Estimators and standard errors of the attack rates are available as a function of p, qE, rE and rC.

Pathogen Evolutionary Genomics to Learn About Transmission

Daniel Wilson

The revolution in DNA sequencing capacity has led to the production of enormous amounts of information concerning the genomic diveristy of pathogens populations. There is great interest in exploiting this source of information to learn about transmission dynamics. Simple approaches compare the evolutionary divergence between pairs of genomes, measured in terms of mutational differences, to impose thresholds that exclude transmission between genomes that are too distantly related. However, this is unlikely to represent the most efficient use of whole pathogen genomes, and does not easily incorporate accompanying epidemiological information. To fully exploit this rich source of data, we need to develop formally integrated evolutionary and epidemiological models that are amenable to likelihood-based inference and based on biologically explicit assumptions. In this talk, I will discuss the idea of using meta-populations to develop such joint evolutionary-epidemiological models, and then apply them to learn about transmission dynamics in Hepatitis C Virus, Staphylococcus aureus and Foot and Mouth Disease Virus. For population-level dynamics, I will review coalescent-based work well suited to analysing genomes sampled at random from large populations. For individual-level dynamics, I will introduce an importance sampling approach that attempts to perform inference under the structured coalescent model. Looking ahead, I will discuss how this latter approach might be generalized to incorporate epidemiological observations, to allow us to reconstruct transmission dynamics using all the information available.

A generalized approach to reconstructing transmission networks for communicable diseases using densely sampled genomic data

Colin Worby

A key aim in the analysis of infectious disease epidemics is to identify who infected whom. Achieving this is challenging, since transmission dynamics are generally unobserved, but a probabilistic estimation of the transmission network based on all available data offers many potential benefits. In particular, this could lead to improved understanding of transmission dynamics, provide a mechanism to quantify factors associated with heightened transmissibility and susceptibility to infection, and help identify effective interventions to reduce transmission. Pathogen typing can assist in the investigation of transmission routes, and whole genome sequence (WGS) data offer maximal discriminatory power, potentially leading to more accurate reconstructions than hitherto possible. Such high resolution genetic data also provide an unprecedented opportunity to investigate the evolutionary behaviour and population dynamics of pathogens. However, uniting the analysis of genetic and surveillance data poses several challenges. In particular, many studies using WGS data have revealed high levels of within-host genetic diversity for common pathogens. To date, genomic data has primarily been used to analyse transmission at a population rather than an individual level. The former typically relies on a broad, low frequency sample of individuals from a large population, with the aim of estimating past population dynamics over a long period of time. In contrast, we focus on individual-level transmission, using high-frequency genomic samples from a subpopulation (eg. hospital, school, jail, farm, community), with the aim of reconstructing transmission routes over a short period of time. We developed a generalized approach to transmission network reconstruction. We made no assumptions about the evolutionary dynamics of the pathogen, and did not consider the phylogenetic relationship between isolates. Instead, we modelled the distribution of genetic distances observed between each pair of sampled isolates. This offers a flexible framework in which multiple independent introductions of the pathogen and within-host diversity may be considered. Each sequenced isolate has both a genetic distance and an epidemiological relationship to each previously observed sequenced isolate. We defined the genetic distance to be the number of SNPs between isolates, though other metrics are possible. The epidemiological relationship describes how the individuals from whom the isolates were taken are linked in the transmission network (for instance, individuals could be part of the same transmission chain, m links apart, or may belong to unrelated transmission chains). This relationship determines the distribution of the genetic distance between isolates intuitively, those who are directly linked would be expected to have isolates differing by fewer SNPs. We used a stochastic transmission model framework, whereby susceptible individuals are exposed to a risk of acquisition, dependent on the number of infectious individuals present in the population at that time. We supposed that each entry to the population was already infected with probability p, and that susceptible persons were homogeneous in terms of risk of acquisition. Let $q(t)$ be the rate at which a given susceptible becomes infected at time t. We

assumed that pathogen detection was not perfect - positive individuals received a positive test result with probability z. Test specificity was assumed to be 100%. A subset of positive isolates was sequenced, and we recorded the genetic distance between each pair of isolates. We suppose that at the time of the sequencing of the nth isolate, n-1 genetic distances are generated. These are drawn from a distribution, which depends on the epidemiological relationship between the n-1 pairs of hosts. One possible model supposes that genetic distances between individuals belonging to separate transmission chains are drawn from a geometric distribution with mean μ_1, while distances between pairs in the same chain, r links apart, are drawn from a geometric distribution with mean $\mu_2\lambda^r$. We primarily used this model, but also experimented with others. If transmission dynamics are perfectly observed, we have a tractable likelihood, and parameter estimation is straightforward. However, since we generally do not know the time, or route, of infection, we augment the parameter space with these missing data, and use a Markov chain Monte Carlo algorithm to explore the posterior distribution. We applied these methods to the transmission of MRSA in hospitals, using data collected from intensive care units in the UK, demonstrating the simultaneous estimation of model parameters and a transmission network. More generally, the approaches we have developed can be applied to the analysis of disease transmission in a community where a high-frequency sample of sequence data is available. These methods offer flexibility not available in previous approaches, such as allowing multiple introductions of the pathogen into the population, incorporating within-host genetic diversity, accounting for uncertainty surrounding colonization times, and providing estimates of uncertainty for each potential transmission route. While we have used whole genome sequence data, this approach may also be used with lower resolution genetic data, provided a distance metric between isolates can be defined.

Estimating transmission trees from genetic data

Rolf Ypma

Knowledge on the transmission tree of an epidemic can provide valuable insights into disease dynamics. The transmission tree can be reconstructed by analysing either detailed epidemiological data (e.g. contact tracing) or, if sufficient genetic diversity accumulates over the course of the epidemic, genetic data of the pathogen. Combining the two disparate data types, genetic and epidemiological data, should yield the best estimates. We present two likelihood-based frameworks we developed to integrate these two data types, estimating probabilities of infection by taking weighted averages over the set of possible transmission trees. In the first, simple framework, we assume conditional independence between all transmission event, using a simplified model of mutation, where substitutions happen at infection rather than at a constant rate over time.[1] We show an application of this approach to a dataset consisting of temporal, geographical and genetic data on the 241 poultry farms infected in an epidemic of avian influenza A (H7N7) in The Netherlands in 2003. The combined approach estimates the transmission

tree with higher correctness and resolution than analyses based on genetic or epidemiological data alone, and the estimated tree reveals the relative infectiousness of farms of different types and sizes. Most remarkably, this approach allows us to get at the actual mechanism of spread between the farms. This mechanism is largely unknown. We tested a putative role of wind, by comparing the direction of estimated transmission events with the direction of wind at that day. We found a statistically significant correlation, which indicates a wind-mediated mechanism of spread.[2] We estimated the contribution of this mechanism to be at least 20% of all infections. In a second, more elaborate, framework, we took into account the full evolutionary history of the sampled sequences. We show how this history is captured by the phylogenetic tree, which itself is contained within the transmission tree.[3] Both the timing and the topology of the two trees can differ. Importantly, the relationship between the two is governed by the within-host viral dynamics. Assuming a simple Wright-Fisher type process, we can jointly infer the transmission tree, phylogenetic tree, within-host dynamics and parameters from genetic and epidemiological data on an outbreak of foot-and-mouth disease. In many instance, the information in the sequences will not be enough to confidently resolve the phylogenetic tree and within-host dynamics, in which case simple approximations such as that in our first framework could be more useful. Approximations further have the nice property of a considerably reduced computation load, mainly due to the conditional independence between transmission events.

References

[1] RJF Ypma et al. *Unravelling transmission trees of infectious disease by combining genetic and epidemiological data*, Proc. R. Soc. B **279(1728)**(2012), 444-50 (IF 5.683).

[2] RJF Ypma et al. *Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza*, Journal of Infectious Diseases **207(5)** (2013), 730-5 (IF 5.848).

[3] RJF Ypma et al. *Relating phylogenetic trees to transmission trees of infectious disease*, outbreaks Genetics 2013 **195(3)** (2013) 1055-62 (IF 4.007).

Reporter: Martin Eichner

Participants

Prof. Dr. Kari Auranen
National Institute for Health & Welfare
Dept. of Vaccination & Immune Protection
P.O. Box 30
00271 Helsinki
FINLAND

Prof. Dr. Frank G. Ball
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Prof. Dr. Niels Becker
National Centre for Epidemiology and Population Health
The Austr. National University
Canberra, ACT 0200
AUSTRALIA

Dr. Martin Bootsma
Department of Mathematics
Utrecht University
P.O.Box 80.010
3508 TA Utrecht
NETHERLANDS

Dr. Tom Britton
Matematiska Institutionen
Stockholms Universitet
10691 Stockholm
SWEDEN

Rosanna Cassidy
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Ben Cooper
Mahidol University
Tropical Medicine Research Unit
60th Anniv., Chalermprakiat Bldg.
420/6 Ratchawithi Rd.
Bangkok 10400
THAILAND

Prof. Dr. Klaus Dietz
Hirschauerstr. 31
72070 Tübingen
GERMANY

Prof. Dr. Martin Eichner
Institut für Medizinische Biometrie
Universität Tübingen
Westbahnhofstr. 55
72070 Tübingen
GERMANY

Prof. Dr. Simon D.W. Frost
Department of Veterinary Medicine
University of Cambridge
Veterinary School
Madingley Rd.
Cambridge CB3 0ES
UNITED KINGDOM

Dr. Sebastian Funk
London School of Hygiene and Tropical Medicine
University of London
Keppel Street
London WC1E 7HT
UNITED KINGDOM

Prof. Dr. Gavin Gibson
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Riccarton
Edinburgh EH14 4AS
UNITED KINGDOM

Dr. Nele Goeyvaerts
Center for Statistics
Hasselt University
Agoralaan Building D
3590 Diepenbeek
BELGIUM

Dr. Gabriela M. Gomes
Instituto Gulbenkian de Ciencia
Apartado 14
Oeiras 2781-901
PORTUGAL

Prof. Dr. M. Elizabeth Halloran
Department of Biostatistics
University of Washington and
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N, LE-400
Seattle, WA 98109-1024
UNITED STATES

Dr. Niel Hens
Center for Statistics
Hasselt University
Agoralaan Building D
3590 Diepenbeek
BELGIUM

Dr. Michael Höhle
Department of Mathematics
Stockholm University
106 91 Stockholm
SWEDEN

Prof. Dr. Edward Ionides
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
UNITED STATES

Prof. Dr. Valerie S. Isham
Dept. of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Niels Keiding
Section of Biostatistics
Kobenhavns Universitet
Oster Farimagsgade 5
P.O. Box 2099
1014 Kobenhavn K
DENMARK

Dr. Eben E. Kenah
Department of Biostatistics
University of Florida
P.O.Box 117450
22 Buckman Dr., 452 Dauer Hall
Gainesville, FL 32610
UNITED STATES

Dr. Mirjam Kretzschmar
Centre for Infectious Disease Control
RIVM
Antonie van Leeuwenhoeklaan 9
P.O.Box 1
3720 BA Bilthoven
NETHERLANDS

Dr. Theodore Kypraios
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Prof. Dr. Chris Leary
Department of Mathematics
SUNY Geneseo
1 College Circle
Geneseo NY 14454
UNITED STATES

Prof. Dr. Ira M. Longini
Department of Biostatistics
University of Florida
P.O.Box 117450
22 Buckman Dr., 452 Dauer Hall
Gainesville, FL 32610
UNITED STATES

Prof. Dr. Emma McBryde
Victorian Infectious Disease Service
Royal Melbourne Hospital
The University of Melbourne
Melbourne, VIC 3010
AUSTRALIA

Dr. Trevelyan J. McKinley
Department of Veterinary Medicine
University of Cambridge
Veterinary School
Madingley Rd.
Cambridge CB3 0ES
UNITED KINGDOM

Prof. Dr. Denis Mollison
The Laigh House
Inveresk
Musselburgh EH21 7TD
UNITED KINGDOM

Dr. Rafal J. Mostowy
Department of Infectious Diseases
Epidemiology
Imperial College London
St. Mary's Campus
Norfolk Place
London W2 1PG
UNITED KINGDOM

Dr. Nico J.D. Nagelkerke
Department of Community Medicine
United Arab Emirates University
P.O.Box 17666
AL Ain
UNITED ARAB EMIRATES

Dr. Peter Neal
Department of Mathematics
University of Lancaster
Fylde College
Bailrigg
Lancaster LA1 4YF
UNITED KINGDOM

Prof. Dr. Philip D. O'Neill
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Lorenzo Pellis
Mathematics Institute
University of Warwick
Zeeman Building
Coventry CV4 7AL
UNITED KINGDOM

David Rasmussen
Department of Biology
Duke University
Box 90338
Durham, NC 27708
UNITED STATES

Prof. Dr. Mick Roberts
Institute of Natural and
Mathematical Sciences
Massey University, Priv. Bag 102904
North Shore City Mail Centre
Auckland
NEW ZEALAND

Eva Santermans
Center for Statistics
Hasselt University
Agoralaan Building D
3590 Diepenbeek
BELGIUM

Prof. Dr. Lisa Sattenspiel
Department of Anthropology
University of Missouri-Columbia
107 Swallow Hall
Columbia MO 65211
UNITED STATES

Dr. Gianpaolo Scalia-Tomba
Dipartimento di Matematica
Universita di Roma Tor Vergata
Via della Ricerca Scientif. 1
00133 Roma
ITALY

Dr. Markus Schwehm
ExploSYS GmbH
Otto-Hahn-Weg 6
70771 Leinfelden
GERMANY

Dr. Simon Spencer
Department of Statistics
University of Warwick
Coventry CV4 7AL
UNITED KINGDOM

Prof. Dr. Nikolaos Stilianakis
Joint Research Centre
European Commission
T.P. 441
Via E. Fermi 1
21027 Ispra (Varese)
ITALY

Dr. Pieter Trapman
Matematiska Institutionen
Stockholms Universitet
10691 Stockholm
SWEDEN

Dr. Michiel van Boven
Centre for Infectious Disease Control
National Institute for Public Health and the Environment
PO Box 1
3720 BA Bilthoven
NETHERLANDS

Prof. Dr. Erik M. Volz
Department of Epidemiology SPH-1
University of Michigan
109 Observatory St.
Ann Arbor MI 48109
UNITED STATES

Prof. Dr. Jon Wakefield
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
UNITED STATES

Dr. Jacco Wallinga
R. I. V. M.
Dept. of Infectious Dis. Epidem.
Antonie van Leeuwenhoeklaan 9
P.O.Box 1
3720 BA Bilthoven
NETHERLANDS

Dr. Daniel Wilson
Wellcome Trust Centre for Human Genetics
University of Oxford
Roosevelt Dr.
Oxford OX3 7BN
UNITED KINGDOM

Dr. Colin Worby
Department of Epidemiology
Harvard School of Public Health
655 Huntington Avenue
Boston MA 02115
UNITED STATES

Dr. Rolf Ypma
Brain Mapping Unit
Department of Psychiatry
Sir William Hardy Building
Downing Street
Cambridge CB2 3EB
UNITED KINGDOM