Mathematisches Forschungsinstitut Oberwolfach

Report No. 8/2023

DOI: 10.4171/OWR/2023/8

Design and Analysis of Infectious Disease Studies

Organized by
Caroline Colijn, Burnaby
M. Elizabeth Halloran, Seattle
Philip O'Neill, Nottingham
Pieter Trapman, Groningen

19 February – 25 February 2023

Abstract. This was the sixth workshop on mathematical and statistical methods for the transmission of infectious diseases. Building on epidemiologic models which were the subject of earlier workshops, this workshop concentrated on disentangling who infected whom by analysing high-resolution genomic data of pathogens which are routinely collected during outbreaks. Following the trail of the small mutations which continuously occur in different places of pathogens' genomes, mathematical tools and computational algorithms were used to reconstruct transmission trees and contact networks. In the past three years these methods were developed and used particularly in the context of the SARS-Cov-2 (Covid-19) pandemic.

Mathematics Subject Classification (2020): 05C05, 05C12, 05C82, 37E25, 37N25, 62F15, 62H30, 62M05, 62M09, 62N01, 62N02, 62P10, 92-04, 92-08, 92C60, 92D20, 92D30.

Introduction by the Organizers

The workshop *Design and Analysis of Infectious Disease Studies*, organized by Caroline Colijn (Burnaby, Canada), M. Elizabeth Halloran (Seattle, USA), Philip O'Neill (Nottingham, UK) and Pieter Trapman (Groningen, the Netherlands), was well attended with 48 participants with broad geographic representation. The participants came from Australia, New Zealand, USA, Brazil, and several countries in Europe, including the UK, Germany, Sweden, Denmark, Finland, Italy, Belgium, and the Netherlands. Thirteen of the 48 participants were women. About ten of the participants were at MFO for the first time. Isaac Goldstein (UC Irvine, USA) and Nicola Mulberry (Simon Fraser University, Canada) were the video conference assistants. One person attended virtually for part of the conference.

Since the last workshop on this topic in 2018, the SARS-Cov-2 (COVID 19) pandemic swept the globe beginning in early 2020. Nearly everyone at this workshop was involved with the response to the pandemic at either the local, national, or international level. There was much discussion among the participants informally about the pandemic. There was a general feeling of elation that this meeting at MFO offered a forum for mathematical, statistical and theoretical discussion free of the daily demands of the pandemic.

The focus of the workshop was on integrating genomic data on pathogens with dynamic epidemiological analysis of infectious disease data either in the endemic or outbreak setting. This is a particularly exciting and challenging area for the analysis of infectious disease data. Now that sequencing the RNA or DNA of viruses, bacteria and other pathogens has become very inexpensive, such data are being obtained from most field studies of infectious diseases. This type of data and evolutionary analysis can contribute a lot to determining who infected whom. Such insight can contribute greatly to public health interventions. The analysis of such data poses statistical, mathematical, theoretical, and computational challenges all at the same time.

There were 24 talks in total, including talks by all five OWLG students. Some of the talks spoke more about the statistical models that were being developed to do such analyses. Other talks dealt with details of computational algorithms. All talks on these related subjects produced active discussions during and after the talks. Other topics included the relevance of social contact patterns for spread of infectious diseases, survival analysis of observational data, and disease burden in transient situations.

As in previous MFO workshops on Design and Analysis of Infectious Disease Studies, there was much discussion in the breaks and in the free periods including consideration of possible collaborations.

The previous workshop in the series has led to several collaborations and publications, e.g.
by Michael Meehan, Daniel G. Cocks, Johannes Müller, Emma S. McBryde (2019) in *J. Math. Biol.*, which was initiated by Müller and McBride at the 2018 workshop and
Xiaoyue Xi, Simon E.F. Spencer, Matthew Hall, M. Kate Grabowski, Joseph Kagaayi and Oliver Ratmann (2022) in *Journal of the Royal Statistical Society: Series C* which was initiated by Spencer and Ratmann at the 2018 workshop.

The 2023 workshop generated many ideas and collaborations as well. Examples include:

- Caroline Colijn and Eben Kenah figured out that a statistical test Colijn proposed for neutral evolution on phylogenetic trees can be rewritten in terms of a score test from a Cox proportional hazards model with a time-dependent exposure. This corrects one problem with the way that the original test handled lineages that go extinct, and it opens up the possibility of controlling for other covariates.

- Oliver Ratmann and Valerie Isham discovered that their mathematical model for estimating time since infection from deep sequence data has an inbuilt bias. At MFO, they developed a better model that should at least in theory resolve the bias. This work builds on early work from the 1980's that Valerie Isham was involved in, and they would not have seen it if it had not been for the great mix of participants at the meeting. Andrea Brizzi (one of the OWLG students) is now taking this model further. This is an important question because they will be using the updated mathematical model to estimate factors associated with late diagnosis of HIV in Africa, a public health question important to UNAIDS and African in-country branches of the Center for Disease Control.
- Niel Hens and Oliver Ratmann sketched out a first version of a statistical model to estimate changes in the distribution of contact patterns that occurred during non pharmaceutical interventions during the COVID-19 pandemic over the past 2 years. This is at present just a very new and attractive mathematical idea, but they believe it should work. It is important because current mathematical models focus on estimating trends in means, whereas in reality one expects time trends in zero-inflation (i.e., a large proportion of individuals stay at home) mixed with other patterns that affect the entire distribution. Hens and Ratmann hope to apply the new model to interpret the data that were collected in Germany and Belgium over the past 2 years and potentially also in the Netherlands and the UK (together this constitutes perhaps the world's largest effort in understanding dynamics in social contacts in the COVID-19 era).
- Claudio Struchiner has initiated conversations with John Edmunds regarding potential collaboration activities. Struchiner invited Edmunds to visit Brazil in August this year (2023) as a first step in strengthening the ties (training activities for students) between the Brazilian and UK institutions.

Most of the participants took the usual Wednesday afternoon hike to St. Roman and back. In St. Roman, we had Schwarzwälderkirschtorte and a choice of beverage.

On Thursday evening, we had our usual musical talent show and cultural event in the lovely music room available at MFO. Dr. Lorenzo Pellis (Manchester, UK) organized the program. The evening's program is presented below.

MFO Talent show, February 23, 2023, 7:30 pm

Lorenzo Pellis (flute) Michiel van Boven (piano)	**Franz Schubert**	Ständchen "Serenade", (adaptation by Van Boven)
Denis Mollison (voice) Kari Auranen (piano)	**Ralph Vaughan Williams**	From far, from eve and morning, words by AE Housman, 2nd song of song cycle "On Wenlock Edge" (1909)
Michiel van Boven (piano)	**Felix Mendelssohn**	Lieder ohne Worte, Opus 30/6
Kari Auranen Michiel van Boven (4-handed piano)	**Johannes Brahms**	Hungarian dance no. 1 in G minor
Caroline Colijn (voice) Kari Auranen (piano)	**Wolfgang Amadeus Mozart**	Laudate Dominum from Vesperae solennes de confessore for solo, choir, orchestra and organ (adaptation by Colijn)
Denis Mollison (voice)	**Denis Mollison**	Poem "Reflections"
Lorenzo Pellis (flute) Michelle Kendall (piano) Simon Spencer (viola)	**Ludwig van Beethoven**	Trio II from Serenade, Opus 25
Caroline Colijn (voice) David Earn (piano)	**Traditional** **Paul Simon**	Siuil a Riun, traditional Irish song Bridge over troubled water
Kari Auranen Elizabeth Halloran (4-handed piano)	**Francis Poulenc**	Finale from Sonata for 4 hands
	Interval	
David Earn (piano)		Improvisation
Lorenzo Pellis (flute) Caroline Colijn (piano)	**Georg Philipp Telemann** **Antonio Vivaldi**	Cantabile from SOLO no. 8 of Essercizzi misiici Largo from Winter from The four seasons
Martin Eichner (voice)	**Sam Walter Foss**	The calf path, originally called "The walk to St. Roman" by Martin Eichner
Kari Auranen (piano)	**Leŏs Janáček**	A blown-away leaf, from On an overgrown path
Michelle Kendall (violin) Lorenzo Pellis (flute)	**Wolfgang Amadeus Mozart** **James Rae**	Duet arrangement of Papageno's Aria "Der Vogelfänger bin ich ja" from Die Zauberflöte Duet no. 2 from Jazzy duets for 2 flutes
Leandro Vendrama (guitar)	**Mariano Mores**	Gricel
Lorenzo Pellis (flute) Caroline Colijn (piano)	**Telemann**	Cantabile from Essercizii Musici
Chris Wymant (voice) Michelle Kendall (piano)	**Sebastian Yatra**	Dos Oroguitas, from the animation film Encanto
Lorenzo Pellis (flute) Elizabeth Halloran Kari Auranen (4 handed piano)	**Georg Friedrich Händel**	Arrival of the Queen of Sheba from Solomon
	Encore	
The bottle blowers	**Traditional** **Bob Dylan**	Frère Jacques Blowin' in the wind
Lorenzo Pellis (flute) Kari Auranen (piano)	**Gaetano Donizetti**	Sonata for flute and piano

Workshop: Design and Analysis of Infectious Disease Studies

Table of Contents

Abstracts

A Bayesian inference method to estimate transmission trees with multiple introductions applied to SARS-CoV-2 in Dutch mink farms

Martin C.J. Bootsma, Don Klinkenberg
(joint work with Bastiaan R. van der Roest, Egil A.J. Fischer, Mirjam E.E. Kretzschmar)

Background

We work on a model and method to infer who infected whom during an infectious disease outbreak, with genetic sequence data. The method applies to the following situation: we have a dataset of an outbreak of an infectious disease that has come to an end. The data of the outbreak consist of the times of detection of all cases, and genetic sequences of the pathogens (e.g. viruses) that infected each host. By using these data, we try to infer who infected whom during the outbreak, and when. We recently developed an extension to the model, that relaxes the assumption that the outbreak started with a single index case. Instead we allow multiple index cases. This extension was developed to analyse a dataset of an outbreak of SARS-CoV-2 in mink farms in the Netherlands [1, 2], that took place in parallel to the epidemic in humans in 2020. In the presentation, we started with an introduction to the method as it was before the extension. Then we presented the model extension for multiple introductions, and discussed some numerical issues and solution related to implementation of the extension in the package phybreak in statistical software R. We finished with results on simulated outbreaks and on the SARS-CoV-2 outbreak in minks.

Methods

In the model we distinguish four submodels for which we can write likelihoods to infer the model parameters and the two variables of interest for each host: when was s/he infected, and by whom. The first is the infection model. In the original model, this starts from one index case, and continues as a branching process of new cases, with each new case having an infector among the existing cases, and an infection time one random generation interval after the infection time of its infector. The second is the detection model: each host is detected (and sampled) one random detection interval after its infection time. The third is the phylogenetic tree model, to create a binary tree describing the 'family' history of the sampled sequences. That is modelled as a coalescent process within each host, conditional on the sampling times and infection times of infectees of that host. This creates phylogenetic minitrees in each host, which are linked through the transmission tree to create one phylogenetic tree. The fourth is the mutation model, conditional on the phylogenetic tree, which creates the genetic variation observed in the sequence data. Mutation is assumed to occur with a Jukes-Cantor mutation model with a single mutation rate, implying that mutations occur with a Poisson process on

the phylogenetic tree, and that each nucleotide change is equally likely. In the extension for multiple introductions, we changed the infection and phylogenetic tree models. In the infection model, instead of assuming a single index case, we now assume a fixed rate by which new index cases can arise after the first index case, until all cases have been detected. These index cases all give rise to their own independent transmission tree. To link the phylogenetic trees of the separate transmission trees, we introduce the concept of a history host, within which the transmission trees are linked through a coalescent process with a rate that is different (lower) than in the observed hosts. The mutation model is the same in observed hosts and the history host.

Implementation

We implemented the extension in an existing package phybreak [3] in statistical software R [4], which infers the transmission and phylogenetic trees with Bayesian MCMC. At first, inferring multiple introductions was problematic, as convergence and mixing of the MCMC chain were slow. We solved slow mixing by implementing an improved mixing algorithm (MC)3 [5], running parallel chains with higher acceptance probabilities to the master chain with the posterior distribution. We solved slow convergence by initializing the MCMC chain with every case as an index case, i.e. infected by the history host, and using the Neighbour-Joining algorithm to create the topology of the phylogenetic tree in the history host.

Simulations

We did a simulation study to test performance of the method. In the simulations of 63 hosts we varied the true number of index cases between 1 and 30, and evaluated how well the method could identify the true number of index cases and the true infector. For the latter, we looked at the posterior most likely infector, and at the posterior 95% set of most likely infectors. It turned out that small numbers of index cases are well identified, but that numbers of index cases may be underestimated when there are many, i.e. roughly more than 10. Across all scenarios with 20 or fewer index cases, the correct infector was identified for 70-75% of all cases, and more than 95% of cases had the true infector in their posterior set of infectors. With 30 index cases, performance was a little worse. Generally two types of errors occurred: incorrect infectors in the correct transmission tree, or merging of transmission trees, thus underestimating the number of index cases.

Results mink farms

When applying our method to the mink farm data, we identified about 13 index cases among 63 farms, 8 more than was concluded from a phylogenetic study on these same data [2]. Because the simulations showed a tendency to underestimate the number of index cases, we are confident that there must have been many more virus introductions than those identified with sequence data alone.

Conclusion and Discussion

We think we developed a useful method to distinguish individual onward transmission from introduction of infections from an outside source, if genetic diversity is limited. A couple of issues is still open for discussion and further development. First, the current method assumes that all cases in the outbreak are observed. This may seem similar to the problem of multiple index cases that we worked on, but it is essentially different because missing cases can be intermediate cases so that the phylogenetic tree can switch between observed and unobserved hosts. The history host in our model is only placed at the root of the outbreak. Missing cases are part of different models and methods, such as the Transphylo model [6], which infers transmission trees on a fixed phylogenetic tree and can thus more easily place unobserved cases between observed cases.

A second point is that in the current model, transmission is described with a branching process that does not take the susceptible population into account. Implicitly, the assumption is made that the population is infinitely large and that there is no depletion of susceptible. In a small population, or in a population with local contact structures (in space or on networks), this is not realistic. To apply the method in small-population settings such as hospital wards, it is necessary to reconsider the transmission model.

A third point concerns the history host. Currently we have assumed a constant-size population in the history host for both the rate of new introductions and the coalescent process. If more information is available, it may be better to use a different population model. For instance, in our own mink farm epidemic we could have used the epidemic curve of the human population. The advantage is that this would have created a more natural rate of new introductions, and an automatic limit of the most-recent common ancestor of the phylogenetic tree within the period of the human epidemic.

References

[1] L. Sikkema, R. Velkers, F. Nieuwenhuijse, D. Fischer, E. .Meijer, et al. *Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands*. Nature Communications, **12** (2021) doi: 10.1038/s41467-021-27096-9.

[2] B. Munnink, R. Sikkema, R. Nieuwenhuijse, D. Molenaar, R. Munger, et al. *Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans*. Science, **371** (2021) 172–7.

[3] D. Klinkenberg, J. Backer, X. Didelot, C. Colijn, and J. Wallinga. *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks*. PLoS Computational Biology **13** (2017) doi: 10.1371/journal.501pcbi.1005495.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*, 2022. URL https://www.r-project.org/

[5] G. ltekar, S. Dwarkadas, J.P. Huelsenbeck, and F. Ronquist. *Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference*. Bioinformatics **20** (2004), 407–15.

[6] X. Didelot, C. Fraser, J. Gardy, C. Colijn, and H. Malik, *Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks*, Molecular Biology and Evolution **34** (2017), 997–1007.

Bayesian inference of timed phylogenetic networks from genomic sequences

Nicola F. Müller

Phylogenetic trees denote the ancestral relationship between sampled individuals. These individuals can be viruses isolated from different hosts, entire species, or event languages. The continual evolution of genomes (or words in the case of languages), means that two individuals that are further apart in their evolutionary history will likely also have more diverged genomes compared to individuals that are more closely related to one another. Using models of evolution and phylogenetic inference techniques, such as Bayesian phylogenetics, allows inferring phylogenetic trees. These phylogenetic trees are formed by population-level processes, such as the transmission of viruses between people and can be inferred from genomic sequences. As such, the shape of these phylogenetic trees, that is the tree topology and branch lengths is different depending on the population process that created them. In turn, we can utilize trees to recover the past population dynamics that created them using phylodynamic models. This allows, for example, to recover past transmission dynamics of pathogens or their global spread from pathogen genomes.

The assumption that the ancestral relationship between sampled individuals can be described by a phylogenetic tree is often necessary to perform inference, at the same time, this assumption is often invalid. For example, upon co-infection of a host, viruses from different lineages can undergo a process of genetic recombination. Different modes of recombination exist, such as reassortment in influenza viruses, the process at the heart of most novel pandemic influenza viruses. In coronaviruses, template switching introduces recombination breakpoints and allows viruses to combine genomes from different ancestral lineages. Recombination occurs frequently in many different coronaviruses, including SARS-like viruses, where recombination is widespread. The same is true for MERS-CoV in camels, the 4 seasonally circulating coronaviruses in humans, as well as for SARS-CoV-2. This poses the question if and how these events are beneficial to those viruses, particularly since these events are relatively widespread.

These processes mean that the ancestral relationship between sampled individuals can no longer be described as a tree, but needs to be described using a phylogenetic network. As such, if we want to learn about how individuals are related in these cases, we need to be able to infer phylogenetic networks from genomic sequences, which requires novel phylogenetic tools.

Here, I present recent work on inferring phylogenetic networks using a Markov chain Monte Carlo approach. First, I describe how the ingredients needed enable inferring phylogenetic networks. These include being able to:

(1) compute network likelihoods, the probability of observing a sequence alignment given a phylogenetic network and model and parameters of sequence evolution.
(2) compute the network prior, the probability of observing a network given a network-generating model and its parameters.

(3) operate on phylogenetic networks to allow exploring the posterior distribution of phylogenetic networks.

1) Simplifies to the product of tree likelihoods. To tackle 2), I present different coalescent bases models that allow modelling reassortment [1], template switching [2], and the horizontal movement of plasmids between bacterial lineages [3]. These coalescent models model a joint coalescent and reassortment, recombination, or plasmid transfer process from present to past where two network lineages can coalesce (share a common ancestor) or reassort (recombine or have plasmid jump between bacteria) [1, 2, 3]. They are parameterized by effective population size (which is inversely proportional to the rate of coalescence) and the rate of reassortment (recombination or plasmid transfer). All of which can be inferred using MCMC. To tackle 3), I touch on the MCMC operations that allow us to relatively efficiently explore different network structures. In particular, I show how we use temporarily augment the space of possible network topologies during MCMC operations to facilitate efficient inference of networks [1]. This allows jointly inferring evolutionary parameters, and phylogenetic networks with the effective population sizes and rates of reassortment (recombination or plasmid transfer) while accounting for uncertainty at each step.

Lastly, I show multiple different applications of phylogenetic network inference. First, I show how we can use the coalescent with reassortment to infer reassortment rates across different influenza viruses, as well as to show how reassortment events are accumulated in parts of the phylogenetic networks that seed future viral variants and therefore that reassortment may contribute to viral fitness [1]. Next, I show how we can use the coalescent with recombination to infer the complex evolutionary history of SARS-like viruses among which are SARS-CoV-1 and 2 [2]. I then show that recombination rates in different coronaviruses, i.e. MERS-CoV-2 and three seasonal human coronaviruses vary between each other with the seasonal human coronaviruses having similar rates of recombination as human influenza viruses reassort, suggesting similar co-infelicitous rates [2]. Lastly, I show how we can use phylogenetic network inference to track the spread of antibiotic resistance gene-carrying plasmids between the different bacterial lineage of Shigella sonnei and Shigella flexneri [3].

References

[1] N.F. Müller, U. Stolz, G. Dudas, T. Stadler and T.G. Vaughan, *Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses*, Proceedings of the National Academy of Sciences **117** (2020), 17104–17111.

[2] N.F. Müller, K.E. Kistler and T. Bedford, *A Bayesian approach to infer recombination patterns in coronaviruses*, Nature communications **13** (2022), 14186.

[3] N.F. Müller, S. Duchêne, D.A. Williamson, B. Howden and D.J. Ingle, *Tracking the horizontal transfer of plasmids in Shigella sonnei and Shigella flexneri using phylogenetics*, bioRxiv (2022), 2022–10.

Recent advances on integrating epidemiological and whole genome sequence data for effectively analysing infectious disease outbreak data

Theodore Kypraios
(joint work with Joseph Marsh, Philip D. O'Neill)

A fundamental aim in the analysis of infectious disease epidemics is to identify who infected whom. However, achieving this is challenging, since transmission dynamics are generally unobserved. A probabilistic estimation of the transmission tree based on all available data offers many potential benefits. In particular, this can lead to improved understanding of transmission dynamics, provide a mechanism to quantify factors associated with heightened transmissibility and susceptibility to carriage and infection, and help identify effective interventions to reduce transmission.

Pathogen typing can be used to cluster genetically similar isolate samples, which can rule out potential transmission routes. Whole genome sequence (WGS) data offers maximal discriminatory power through the identification of individual point mutations, or single nucleotide polymorphisms (SNPs), potentially leading to more accurate transmission tree reconstructions than hitherto possible. However, the joint analysis of genetic and surveillance data poses several challenges, as the relationship between epidemic and evolutionary dynamics is complex.

Despite recent advances in statistical models and methods for combining genomic data with traditional epidemiological data (e.g. incidence data), existing approaches have their own limitations, such as simplifications to the underlying biological processes, arbitrary phenomenological models or approximations to the likelihood function, to name a few.

We focus on individual-level transmission, using genomic samples from a subpopulation (eg. hospital, school, jail, farm, community), with the aim of reconstructing transmission routes. Following our earlier work [1, 2], we present a modelling framework for integrating epidemiological and whole genome sequence data where we use the matrix of pairwise horizontal distances between sequences as a summary statistic for the genetic data. We address the limitations of existing approaches, in which these distances are modelled using phenomenological models by explicitly deriving the joint probability distribution of pairwise genetic distances under the assumption of a microevolution mutation model.

Our proposed framework allows data to be simulated forward in time, a feature lacking in the majority of existing methods (with reverse time simulation typically required in phylogenetic methods, and only an incomplete set of genetic distances simulated from other approaches), which is of fundamental importance in predictive modelling and model evaluation. We provide a framework with the flexibility to allow for unobserved infection times, multiple independent introductions of the pathogen, and within-host genetic diversity.

We develop bespoke data-augmentation MCMC algorithms to infer the transmission network and the unobserved pathogen distances at the time of transmission as well as the times of transmission. We illustrate the predictive performance

of our methodology using simulated data, as well as analysing data from an an outbreak of *S. aureus* in an intensive care unit in Brighton during 2011-2012.

References

[1] C.J. Worby, P.D. O'Neill, T. Kypraios, J.V. Robotham, D. De Angelis, E.J. Cartwright, S.J. Peacock, and B.S. Cooper, *Reconstructing transmission trees for communicable diseases using densely sampled genetic data*. The Annals of Applied Statistics, **10**:1 (2016), 395–417

[2] R. Cassidy, T. Kypraios, and P.D. O'Neill, *Modelling, Bayesian inference, and model assessment for nosocomial pathogens using whole–genome–sequence data*. Statistics in Medicine, **39**:12 (2020), 1746–1765.

How did behavioural patterns, seasonality and virus strains affect transmission during Covid-19? and Optimal intervention strategies for minimizing total incidence during an epidemic

Tom Britton

(joint work with Felix Gunther, Hilde Kjelgaard Brustad, Arnoldo Frigessi, Lasse Leskela)

The first part concerns a project where we try to make use of various time-dependent data sources on: temperature, virus strain frequencies, behaviour metrics recorded by google, and vaccination data, to try to infer how these factors affected the spread of Covid-19. Our focus is on Norway and Sweden, and the during the year 2021.

Using computer intensive Bayesian methodology we analyse an epidemic model on a regional level to infer how much of the observed transmission that can be attributed to various behaviour, and how virus strains and temperature (seasonality) affect transmission dynamics.

Some main findings are that the four google metrics "Work", "Transit", "Grocery and pharmacies" and "Retail and recreation" capture 40-70% of all transmission, and the former two, "Work" and "Transit", have a bigger impact on transmission than "Grocery and pharmacies" and "Retail and recreation". The remaining transmission taking place is captured in nuisance parameters and may for example happen in the households but also in various other settings. The google metric "Household", was left out in the analysis because there was a very strong negative correlation between this behaviour and all the other factors. Further, the temperature/seasonality effect imply that transmission is reduced by 40-50% in summer as compared to winter season.

The estimates of the different regions as well as between the two countries are fairly robust. The analysis can be used to learn more regarding which behavioural changes are most important for reducing transmission should a new pandemic arrive, and also to estimate how much different behaviours need to change in order to reduce the reproduction number below the critical value 1.

This is work in progress.

The second part is purely mathematical/theoretical. Here the scientific questions is how to optimally intervene in an epidemic if the aim is to minimize the total fraction τ getting infected (or equivalently the total number of hospitalisations or case fatalities). More specifically, we study the simple deterministic SIR epidemic and assume that the rate of infectious contact β may be reduced by a factor $p(t)$ at time t. The optimization problem is then to find which preventive strategy $\{p(t), 0 \leq t < \infty$, such that the overall amount of prevention $\int_0^\infty p(t)dt \leq c$ is no bigger than some given maximum prevention cost c, minimizes the total fraction τ getting infected. It is worth pointing out that the space of possible prevention strategies is very large: full lockdown ($p(t) = 1$) for c days over some specified interval, half-lockdown during $2c$ days, linearly growing/declining lockdown, separate days/weeks of full/partial lockdown, and so on.

Using optimization and properties of the SIR epidemic model we are able to prove that that optimal is simple and also quite surprising. More specifically, it consists of doing nothing until the unrestricted epidemic peaks, and then insert a maximal lockdown for c days. If it is not possible to reduce transmission more than some maximum value p_{max} (so $p(t) \leq p_{max}$ for all t), then the optimal solution is very similar: do nothing up until close to the peak (how close depends on p_{max} and then insert a maximal lockdown $p(t) = p_{max}$ for c/p_{max} days and then go back to normality.

It is also shown that the effect of adding (e.g. a week) lockdown before the peak of unrestricted peak actually *increases* τ, the total fraction getting infected!

This is joint work with Lasse Leskela [1]

References

[1] T. Britton and L. Leskela, *Optimal intervention strategies for minimizing total incidence during an epidemic*,SIAM Journal on Applied Mathematics **83** (2023), 354–373.

Doublethink: Identifying epidemiological risk factors in UK Biobank using simultaneous Bayesian/frequentist model averaging

Daniel Wilson

Epidemiological discovery of risk factors using linear models is a variable selection problem. Correlation (non-orthogonality) between candidate risk factors creates a challenge for interpretation: the evidence that variable A is a risk factor often changes depending on whether correlated variable B is included or not. However, systematic variable selection is rarely pursued in practice because of (a) disinclination toward Bayesian methods, including (b) the need to use subjective prior distributions instead of objective frequentist false positive rates, (c) computational inviability of exhaustive search in big data settings and (d) computational intensity of Monte Carlo methods. To address these challenges, we have developed Doublethink for simultaneous Bayesian/frequentist inference of risk factors via model averaging. Doublethink facilitates interchangeable thresholding between false positive rates and posterior probabilities, and ameliorates the computational

load through a large-sample-motivated approximate likelihood. We applied the method to identify risk factors for COVID-19 outcomes (infection, hospitalization, mortality) among $> 100,000$ UK Biobank participants and $> 1,000$ candidate risk factors. We compare and contrast our results to the literature. We find that this systematic approach produces less sparse accounts of risk factor importance, underlining the importance of many lifestyle, medical and environmental phenomena on infection and disease outcomes.

Exact Phylodynamic Likelihood

Aaron A. King
(joint work with Qianying Lin, Edward L. Ionides)

The project of phylodynamics is the extraction of information on the nature of a stochastic population process from data on the relationships among genomic samples taken from individuals in the evolving population. In its purest form, its core problem may be factored into two subproblems: the identification of genealogical trees expressing the relationships between genomic samples and the probabilistic linkage of these genealogies to the generating stochastic process. In this work, we focus on the latter. Specifically, we view the genealogy as data and aim to compute the likelihood of the genealogy under any given Markovian population process.

Two distinct approaches to this problem already exist. The older builds on the Kingman coalescent [1, 2, 3] and variations thereon; the younger, on generalized linear birth-death processes [6, 7, 5]. The key element in the tractability of both approaches has been the computability certain approximate reverse-time transition probabilities, but these approximations are only accurate in the limit of large population size and/or small sample fraction. In this work, we aim to eliminate the need for such approximations.

To accomplish this, we construct a novel class of genealogy-valued Markov processes, each uniquely induced by any given discretely-structured Markovian population process. The latter class is sufficiently rich as to encompass most infectious-disease transmission models of practical interest. Preliminary results for the unstructured case were given in [4]. We present a theorem giving the exact probability distribution of genealogies conditional on the history of the population process. We then show how integration over the space of population histories yields a nonlinear filtering equation with continuous and discrete portions. This equation may be integrated via well understood Feynman-Kaç approaches, which take the form of certain sequential Monte Carlo algorithms.

The results are a strict generalization and unification of existing approaches. The proofs rely on several constructions which are both novel and more natural than the reverse-time constructions used in the more limited coalescent- and birth-death-process-based theories. Importantly for applications, the implied algorithms can be carried out entirely in forward time.

References

[1] J. F. C. Kingman, *The coalescent*, Stochastic Processes and their Applications **13** (1982), 235–248.

[2] J. F. C. Kingman, *Exchangeability and the evolution of large populations*, in G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112, North-Holland, Amsterdam (1982).

[3] J. F. C. Kingman, *On the genealogy of large populations*, Journal of Applied Probability **19** (1982), 27–43.

[4] A. A. King, Q. Lin, and E. L. Ionides, *Markov genealogy processes*, Theoretical Population Biology **143** (2022), 77–91.

[5] N. F. Müller, D. A. Rasmussen, and T. Stadler, *The structured coalescent and its approximations*, Molecular Biology and Evolution **34** (2017), 2970–2981.

[6] T. Stadler, *Sampling-through-time in birth-death trees*, Journal of Theoretical Biology **267** (2010), 396–404.

[7] T. Stadler, R. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, H. F. Günthard, A. J. Drummond, S. Bonhoeffer, and the Swiss H.I.V. Cohort Study, *Estimating the basic reproductive number from viral sequence data*, Molecular Biology and Evolution **29** (2012), 347–357.

Does branching now imply branching next? Testing for exchangeability in timed phylogenies and Estimating serial intervals with genomic data

Caroline Colijn

(joint work with Jessica E. Stockdale, Kurnia Susvitasari, Paul Tupper, Benjamin Sobkowiak, Nicola Mulberry, Anders Gonçalves da Silva, Anne E. Watt, Norelle Sherry, Corinna Minko, Benjamin P. Howden, Courtney R. Lane)

In part 1 of the talk, titled **Does branching now imply branching next? Testing for exchangeability in timed phylogenies**, by C. Colijn (sole author), I develop a statistical test for exchangeability on timed phylogenies. The test is based on nodes in a phylogeny for which one of their child lineages branches next after them. Under the null hypothesis (exchangeability), the probability that this occurs is m/k, where m is the number of child lineages and k is the number of lineages in total at the node's time. I write a test statistic using these probabilities, and do hypothesis testing to test for exchangeability. I illustrate this with simulated data and with sequence data from SARS-CoV-2.

In the second part of the talk, titled **Estimating serial intervals with genomic data**, with the full listed authorship group, I present a method that uses pathogen genomes and symptom onset times to infer the serial interval distribution in a cluster or outbreak. The serial interval is the time between symptom onset in a pair A and B, where A infected B; it is a key parameter for infectious disease modelling. To use genomes to estimate this, we must take uncertainty in the who-infected-whom pairs into account; we sample many possible transmission trees, and use a mixture model to account for ways that missing individuals might occur. Our method can estimate the distribution from data that are often routinely

collected. This is in contrast to standard methods, which require labour-intensive and private data.

Likelihood-based Inference for Stochastic Epidemic Models via Data Augmentation

Jason Xu

Stochastic epidemic models such as the Susceptible-Infectious-Removed (SIR) model are widely used to model the spread of disease at the population level, but fitting these models present significant challenges when missing data or latent variables are present. In particular, the likelihood function of the partially observed data is typically considered intractable in many common observational settings, such as when incidence data on new counts are collected. We will discuss recent advances that enable likelihood computations without model simplifications in the presence of missing infection and recovery times via efficient data-augmented samplers. Our methods target the exact posterior without relying on model-based forward simulation, and apply to several classic stochastic compartmental models and allow for disease-dependent contact networks to evolve dynamically.

This is part of a broader research effort to revisit classical Bayesian Markov chain Monte Carlo (MCMC) sampling schemes, such as Metropolis-Hastings and Metropolis-within-Gibbs, to enable *likelihood-based* for nonlinear stochastic models of disease. Arguably, the predominant approach for fully stochastic models entails simulation methods, which become computationally costly and can face degeneracy issues for some data settings. Our approach complements this well-established approach, combining new ideas for designing efficient proposal densities with modern computational statistics to revive a classic line of thought on direct likelihood-based inference for such models. In the past, we presented a generating function inversion approach for integrating over the space of possible configurations of unobserved quantities, accounting for their probabilistic paths exactly and thereby yielding the marginal data likelihood. In a classic text Stochastic Population Processes, Renshaw [1] remarks that "the associated mathematical manipulations required to generate solutions can only be described as heroic". We recently resolved this problem in a class of bivariate competition processes that includes the SIR model. The method is delicate, however, relying on continued fraction expansions in the Laplace domain of transition probabilities that do not carry over easily to generalizations of the model.

Our current work instead focuses on enabling this kind of marginalization through latent variables in a flexible MCMC framework. We make use of a powerful data augmentation strategy that carefully proposes high-dimensional latent variables to efficiently explore the possible configurations of unobservable quantities in nonlinear stochastic models. This opens the door to efficient yet flexible Markov chain Monte Carlo methods, and allow us to relax stringent classical assumptions. In the context of epidemics, it enables us to model the co-evolution of the disease process together with a dynamic network describing the contact

patterns, replacing a stringent well-mixing assumption. The results presented at this workshop show how the methodology enabled new insights on a mobile health study of flu-like illness on a college campus.

We also derived extensions to the semi-Markov setting, enabling more realistic renewal distributions for disease latency and time until recovery. The data augmented framework makes these extensions possible immediately, and we discuss high-level intuition for the core idea of decoupling non-linear dynamics into tractable, simpler surrogate processes. By generating proposals from classes such as multitype branching processes, whose properties enable efficient *conditional* simulation to propose trajectories consistent with observed data, we can take large steps in the latent space. As a result, Markov chain samplers can efficiently explore the configurations of unobserved variables, avoiding the practical limitations that lead existing approaches to poor mixing and prohibitively high autocorrelation in the samples. The methods we propose enable us to perform exact inference on large outbreaks even using just a single laptop. Many future directions remain open in making the modelling framework more realistic and extensible, incorporating various sources of heterogeneity and population stratification.

References

[1] E. Renshaw, *Stochastic Population Processes: Analysis, Approximations, Simulations*, Oxford University Press (2011).

Multi-strain models for nosocomial infections

Alice Thompson
(joint work with Philip O'Neill and Theodore Kypraios)

Nosocomial infections have been a growing problem in hospitals for over 50 years, with most of these infections being due to the introduction and misuse of antibiotics. With motivation from a unique data set, this talk discusses how transmission networks can be reconstructed for a multi-strain outbreak in a hospital environment. This research focuses on the use of genetic data, collected using Whole Genome Sequencing techniques, and epidemiological data from patients and surfaces to recon- struct transmission networks of various stains and species of Klebsiella pneumoniae Carbapenemase (KPC)-producing bacteria. In this talk, we will first introduce the contents of the data set. This includes a discussion regarding the quantity of both patient and environmental test data and the diversity of samples sequenced. Then, we discuss the construction of the multi-strain transmission model and its augmented likelihood. Next, we discuss the use of MCMC and data augmentation which enables us to estimate parameters and reconstruct multiple transmission networks using both genetic and epidemiological data, and the results yielded when using this method on a simulated data set. We will show the accuracy of the methods parameter estimates and its successful convergence, along with applying the method to a data set similar to our motivational data set previously introduced. Finally, a new environmental model will be introduced which

also incorporates environmental data and enables envi- ronment to be part of the transmission network. We also introduce two approaches on handling cleaning on wards.

Phylogenetic Estimation of HIV Time Since Infection

Andrea Brizzi
(joint work with Oliver Ratmann)

The distribution of the delay from HIV infection to diagnosis is of primary epidemiological interest, as it can inform changes in testing policy. However, determining the date of infection of a person living with HIV is often hard, due the typical asymptomatic period lasting years.

In middle- and high-income countries, the dating of infection is therefore generally based on the modelling of longitudinal biomarker measurements, such as CD4 counts and viral load. Analyses based on these biomarker methods have been used to determine the place of HIV acquisition for migrants in Europe [1] or to determine differences in delay to diagnosis among ethnic groups in London [2].

In lower-income countries, where longitudinal studies may be too complicated or expensive to carry out, phylogenetic methods can help extract as much historic information as possible from a single blood sample. For example, the Random Forest algorithm "HIV-phyloTSI" [3], uses features obtained from phylogenies as predictors to estimate the delay from infection to sample collection at the population level. We coded a pipeline available at `https://github.com/olli0601/Phyloscanner.R.utilities/tree/master/misc_data_analysis_RCCS1519/software` and obtained predictions for 7103 sequences from the Rakai Community Cohort Study in Uganda.

We then selected for further analysis the subset of 5170 sequences sampled at most 3 months after diagnosis. We run a Bayesian model to obtain smoothed group-level estimates from the individual-level estimates obtained from phylo-TSI. In particular, the individual-level estimates were assumed to follow a Gamma distribution, and the mean delay to diagnosis was modelled through gender and cohort specific baselines in to a sex-specific random functions on age at diagnosis and a random function on year of diagnosis. The model was fit via Hamiltonian Monte Carlo through Stan and the random functions were modelled as Hilbert Space Gaussian Process approximations.

The results highlight that delay to diagnosis increases with age at diagnosis. Although this may seem intuitive, it is interesting to note that the HIV-phylo-TSI algorithm does not use age of the host as a predictor, and therefore this results is purely obtained from the phylogenetic signal.

However, the model does not seem to find significant differences in delay to diagnoses in men and women, or reductions in delay to diagnoses over time, which are both hypothesised to play a role in the Rakai settings.

Further work will try to more rigorously test these differences by including the limited biomarker data available and including a survival component to the Bayesian model.

References

[1] N. Pantazis et. al. *Determining the likely place of HIV acquisition for migrants in Europe combining subject-specific information and biomarkers data*, Statistical Methods in Medical Research, **28(7)**, (2017), 1979–1997.

[2] O. Stirrup and D. Dunn *Estimation of delay to diagnosis and incidence in HIV using indirect evidence of infection dates*, BMC Medical research Methodology, **18(1)** (2018).

[3] T. Golubchik et. al. *HIV-phyloTSI: Subtype-independent estimation of time since HIV-1 infection for cross-sectional measures of population incidence using deep sequence data*, MedArxiv

A nested model for pneumococcal population dynamics

Nicola Mulberry

(joint work with Alexander Rutherford, Caroline Colijn)

Streptococcus pneumoniae is a pathogen of major public health concern globally. Pneumococcal strains exhibit diversity in their capsular serotype, metabolic profiles, and properties of antibiotic resistance (among other traits). Pneumococcal conjugate vaccines have been successful at targeting a subset of the circulating serotypes. Following such perturbation, pneumococcal populations have been shown to undergo significant shifts indicative of competition both between and within serotypes. Using a nested model with explicit within-host dynamics, we show how competition between these types, along with heterogeneity in duration of carriage, may help explain patterns of vaccine-induced population dynamics.

Epidemic models with manual and digital contact tracing

Dongni Zhang

(joint work with Tom Britton)

We consider a Markovian SIR epidemic model in a homogeneous mixing community with a constant rate of diagnosis (testing) and investigate the preventive effects of two types of contact tracing (CT): manual and digital CT.

In [1], we introduce the traditional manual CT by assuming that once an infectious individual tests positive, s/he is immediately isolated and each of her/his contacts are traced and tested independently with some fixed probability. Using large population approximations, we analysed the early stage of the outbreak when the process of "to-be-traced components" behaves like a branching process. The component and individual reproduction numbers are derived. In [2], we focus on the more recent digital CT via a tracing app (only app-users can trigger and be traced by digital tracing). We assume that manual or digital CT occurs instantaneously and recursively for mathematical tractability. The model with digital CT is analysed by a two-type branching process relying on a large community,

where one type of "individuals" are "app-using components" and another is non-app-users. Further, we investigate the combined preventive effect of manual and digital CT. This combined model is analysed by a different two-type branching process with both types being the "to-be-traced components" but starting with different "roots". The corresponding reproduction numbers are derived. We conclude that it is more essential to control the epidemic to have a large fraction of app-users compared to the manual tracing probability. Another important conclusion is that the combined effect is bigger than the product of two separate preventive effects.

The ongoing work is to generalize the combined model above by first assuming that it is not only possible to infect neighbours in a network (e.g. daily/recent close contacts), but also transmission could happen from random type of contacts (neighbour or not, e.g. on a bus), which are usually more easily and quickly identified by using a tracing app; and then incorporate the manual CT only on the network with tracing delay but the instantaneous digital CT both on network and among global contacts.

References

[1] D. Zhang and T. Britton, *Analysing the Effect of Test-and-Trace Strategy in an SIR Epidemic Model*, Bulletin of Mathematical Biology **84**(10) (2022), 105.

[2] D. Zhang and T. Britton, *Epidemic models with digital and manual contact tracing*, arXiv preprint arXiv:2211.12869 (2022).

Flow models to interpret population-based deep- sequence pathogen data, with application to longitudinal sequence and surveillance data from East Africa

Oliver Ratmann

(joint work with Melodie Monod, Andrea Brizzi, Alexandra Blenkinsop, Yu Chen and Shozen Dan)

HIV incidence in eastern and southern Africa has historically been concentrated among girls and women aged 15-24 years, but as new cases decline with HIV interventions, population-level infection dynamics may shift by age and gender. Our mathematical work is concerned with the problem of quantifying how HIV incidence and the population groups driving transmission have evolved over a 15 year period from 2003 to 2018 in Uganda, based on population-based surveillance and longitudinal deep-sequence data. The Rakai Community Cohort Study (RCCS) encompasses both a full census of the study communities and a population-based survey in each surveillance round, which enables identification and follow up of unique individuals over time, and thus provides a comprehensive sampling frame to measure HIV incidence. HIV incidence was estimated using established GAMLSS inference methods with the `R` package `mgcv`. Rather exceptionally, the RCCS also performed population-based HIV deep-sequencing spanning a period of more than 6 years, from August 2011 to April 2018. The primary purpose of viral deep

sequencing was to reconstruct transmission networks and identify the population-level sources of infections, thus complementing the data collected through the incidence cohort. The RCCS viral phylogenetic transmission cohort comprises of all participants with HIV for whom at least one HIV deep sequence sample satisfying minimum quality criteria for deep-sequence phylogenetic analysis is available. The HIV deep-sequencing pipeline developed by PANGEA-HIV then provided sequence fragments that capture viral diversity within individuals, which enables phylogenetic inference into the direction of transmission from sequence data alone with the `phyloscanner` software. First, potential transmission networks were identified, and in the second step transmission networks were confirmed and the transmission directions in the networks were characterised as possible. We here developed a framework for estimating the sources of the population-level HIV incidence dynamics from the dated, source-recipient pairs in the viral phylogenetic transmission cohort. Overall, inference was done in a Bayesian framework using a semi-parametric Poisson flow model similar to Xi, X. *et al.* [3], that was fitted to observed counts of transmission flows $Y_{p,i,j}^{g\to h}$ with transmission direction $g \to h$ (male-to-female or female-to-male), time period p (R10-R15 and R16-R18) in which the recipient was likely infected, and 1-year age bands i, j of the source and recipient populations respectively, where

$$i, j \in \mathcal{A} = \{15, 16, \ldots, 48, 49\} \tag{1a}$$

$$(g \to h) \in \mathcal{D} = \{\text{male-to-female}, \text{female-to-male}\}. \tag{1b}$$

The target quantity of the model is the expected number of HIV transmissions in the study population in transmission direction $g \to h$ (male-to-female or female-to-male), survey round r (R10 to R18) in which infection occurred, and 1-year age bands i, j of the source and recipient populations respectively, which we denote by $\lambda_{r,i,j}^{g\to h}$. We considered that the expected number of HIV transmissions in the study population is characterized by transmission risk and modulated by the number of infectious and susceptible individuals, which prompted us to express $\lambda_{r,i,j}^{g\to h}$ in the form of a standard discrete-time susceptible-infected (SI) model,

$$\lambda_{r,i,j}^{g\to h} = \beta_{r,i,j}^{g\to h} \times S_{r,j}^{h} \times I_{r,i}^{g} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right|, \tag{2}$$

where $\beta_{r,i,j}^{g\to h} > 0$ is the transmission rate exerted by one infected, virally unsuppressed individual of gender g and age i on one person in the uninfected ("susceptible") population of the opposite gender h and age j in a standardized unit of time in round r. With model (2), we express expected transmission flows with a population-level mechanism of how transmission rates from individuals with unsuppressed HIV act on the susceptible population, and we preferred model (2) over a purely phenomenological model of the $\lambda_{r,i,j}^{g\to h}$ for the generalizing insights it provides. The main simplifying approximations in (2) are that all quantities on the right-hand side of (2) are in discrete time and constant in each round, meaning we approximate over changes in population size, HIV prevalence, and viral suppression at a temporally finer scale, and assume further that one generation of

transmissions occurs from individuals with unsuppressed HIV in each round. Importantly, in this framework, we can then relate the expected transmission flows to the HIV incidence dynamics and the data from the longitudinal incidence cohort by summing in (2) over the sources of infections,

$$\sum_i \lambda_{r,i,j}^{g\to h} = \left(\sum_i \beta_{r,i,j}^{g\to h} \times I_{r,i}^{g}\right) \times S_{r,j}^{h} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right| \tag{3a}$$

$$=: \kappa_{r,j}^{h} \times S_{r,j}^{h} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right|, \tag{3b}$$

where $\kappa_{r,j}^{h}$ is the incidence rate per census-eligible, susceptible person of gender h and age j in round r ($S_{r,j}^{h}$) and per unit time ($|(t_r^{\text{end}} - t_r^{\text{start}})|$). Estimates of $\kappa_{r,j}^{h}$ were calculated in units of 100 person-years, and we will constrain the semi-parametric Poisson flow model using these estimates. From the model output, we are primarily interested in the transmission flows and transmission sources during each round as quantities out of 100%, defined respectively by

$$\pi_{r,i,j}^{g\to h} = \lambda_{r,i,j}^{g\to h} \Big/ \Big(\sum_{i,j\in\mathcal{A},(g\to h)\in\mathcal{D}} \lambda_{r,i,j}^{g\to h}\Big) \tag{4a}$$

$$\delta_{r,i,j}^{g\to h} = \pi_{r,i,j}^{g\to h} \Big/ \Big(\sum_{k\in\mathcal{A}} \pi_{r,k,j}^{g\to h}\Big) \tag{4b}$$

$$\delta_{r,i}^{g\to h} = \sum_{j\in\mathcal{A}} \pi_{r,i,j}^{g\to h}\,. \tag{4c}$$

In words, (4b) quantifies the sources of infection in individuals of gender h and age j in round r such that the sum of $\delta_{r,i,j}^{g\to h}$ over i equals one, and (4c) quantifies the sources of infection in the entire population in round r that originate from the group of individuals of gender g and age i such that the sum of $\delta_{r,i}^{g\to h}$ over g and i equals one. The number $S_{r,j}^{h}$ of the susceptible population of gender h and age j was calculated by multiplying the smoothed estimate $N_{r,j}^{g}$ of the census-eligible population of gender h and age j with 1 minus the posterior median estimate of HIV prevalence $\rho_{r,j}^{h}$ in census-eligible individuals of gender h and age j of round r (calculated as described further above). To specify the number $I_{r,i}^{g}$ of individuals with unsuppressed HIV of gender g and age i, we multiplied the smoothed estimate $N_{r,i}^{g}$ of the census-eligible population of gender g and age i of round r with the posterior median estimate of HIV prevalence in the census-eligible population of gender g and age i ($\rho_{r,i}^{g}$) with 1 minus the posterior median estimate $\nu_{r,i}^{g}$ of the proportion of census-eligible individuals of gender g and age i in round r that have suppressed HIV. We first present the likelihood of the observed counts of transmission flows $Y_{p,i,j}^{g\to h}$ under the semi-parametric Poisson flow model that is parameterised in terms of (2). The phylogenetically reconstructed source-recipient pairs capture only a subset of incidence events, and so it is important to characterise the sampling frame. Because we are here integrating data from the transmission and incidence cohorts, we are able to adjust inferences by detection probabilities of incidence events. Specifically, we express the detection probability as the ratio of phylogenetically reconstructed transmission events with a recipient

of gender h and age j divided by the expected number of incident cases of gender h and age j in time period p as derived in (3),

$$\xi^h_{p,j} = \Big(\sum_{i\in\mathcal{A}} Y^{g\to h}_{p,i,j}\Big) \Big/ \Big(\sum_{r\in p} \kappa^h_{r,j} \times S^h_{r,j} \times |(t^{\text{end}}_r - t^{\text{start}}_r)|\Big). \tag{5}$$

We assume in (5) that the detection probability does not depend on characteristics of the source, further characteristics of the recipient beyond their age and gender, and is constant in time period p. These assumptions imply that infection events are sampled identically and independently with probability (5), which in turn allows us to express the likelihood of observing the phylogenetic data similarly as in Xi, X. *et al.* [3] with

$$Y^{g\to h}_{p,i,j} \sim \text{Poisson}\Big(\xi^h_{p,j} \sum_{r\in p} \lambda^{g\to h}_{r,i,j}\Big) \tag{6a}$$

$$\lambda^{g\to h}_{r,i,j} = \beta^{g\to h}_{r,i,j} \times S^h_{r,j} \times I^g_{r,i} \times \big|\big(t^{\text{end}}_r - t^{\text{start}}_r\big)\big| \tag{6b}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(i), \tag{6c}$$

where $\hat{\boldsymbol{c}}^{g\to h}(i,j)$ is the posterior median estimate of the log rate of sexual contacts within communities in one year between one person of age i and gender g and one person of age j and gender h that we estimated from the sexual behaviour data, and the remaining terms quantify the transmission probability per sexual contact on the log scale. The model is designed in such a way that the log sexual contact rates describe a fixed age-specific non-zero mean surface, and the remaining parameters describe age-specific random deviations around the mean surface. With this approach, any inferred deviations in transmission rates relative to sexual contact rates are informed by the phylogenetic data and robust to prior specifications on the random deviations. Specifically, γ_0 is the baseline parameter characterising overall transmission risk per sexual contact, γ_g is a gender-specific offset which is set to zero in the female-to-male direction and a real value in male-to-female direction, γ_r a round-specific offset which is set to zero for the first survey round 10, and γ_p is a time period specific offset which is set to zero for the first time period. We assume the age-specific structure of transmission rates in terms of the transmitting partners (denoted by i) and recipients (denoted by j) are similar across similar ages, and so we can exploit regularising prior densities [3] to learn smooth, latent transmission rate surfaces from the sparse data. In detail, we modelled the age-specific structure of transmission rates non-parametrically with 2 time-invariant random functions $\boldsymbol{f}^{g\to h}_0$ with two-dimensional inputs on the domain $[15, 50] \times [15, 50]$ that characterise age-age interactions in transmission risk for each gender, 2×8 random functions $\boldsymbol{f}^{g\to h}_r$ with one-dimensional inputs that characterise time trends in the age of recipients for each gender for survey rounds after round 10, and 2 random functions $\boldsymbol{f}^{g\to h}_p$ with one-dimensional inputs that characterise time trends in the age of transmitting partners for each gender for the

second time period. We attach to each of these random functions computationally efficient B-splines projected Gaussian process (GP) priors [2], which we constructed by describing the random functions with cubic B-splines over equidistant knots and modelling the prior relationship of the B-splines parameters with GPs with squared exponential kernels with variance and lengthscale hyper-parameters, denoted respectively by σ^2 and ℓ. The prior densities of our Bayesian model are

$$\begin{aligned}
&\text{(7a)} && \gamma_0 \sim \mathcal{N}(0, 10^2) && \\
&\text{(7b)} && \gamma_{\text{male}} \sim \mathcal{N}(0, 1) && \\
&\text{(7c)} && \gamma_r \sim \mathcal{N}(0, 1) && \text{for } r > \text{R10} \\
&\text{(7d)} && \gamma_p \sim \mathcal{N}(0, 1) && \text{for } p = \text{R16-R18} \\
&\text{(7e)} && \boldsymbol{f}_0^{g\to h} \sim \text{2D-B-splines-GP}(\sigma_0^{g\to h}, \ell_{0,i}^{g\to h}, \ell_{0,j}^{g\to h}) && \\
&\text{(7f)} && \boldsymbol{f}_r^{g\to h} \sim \text{1D-B-splines-GP}(\tilde{\sigma}_r^{g\to h}, \tilde{\ell}_r^{g\to h}) && \text{for } r > \text{R10} \\
&\text{(7g)} && \boldsymbol{f}_p^{g\to h} \sim \text{1D-B-splines-GP}(\breve{\sigma}^{g\to h}, \breve{\ell}^{g\to h}) && \text{for } p = \text{R16-R18} \\
&\text{(7h)} && \sigma_{0,i}^{g\to h}, \sigma_{0,j}^{g\to h}, \tilde{\sigma}^{g\to h}, \breve{\sigma}^{g\to h} \sim \text{Half-Cauchy}(0, 1) && \\
&\text{(7i)} && \ell_{0,i}^{g\to h}, \ell_{0,j}^{g\to h}, \tilde{\ell}^{g\to h}, \breve{\ell}^{g\to h} \sim \text{Inv-Gamma}(2, 2), &&
\end{aligned}$$

where the 2×8 recipient-specific time-varying 1D B-splines GPs each have squared exponential kernels with hyper-parameters $\tilde{\sigma}_r^{g\to h}$, $\tilde{\ell}^{g\to h}$, the 2 source-specific time-varying 1D B-splines GPs each have squared exponential kernels with hyper-parameters $\breve{\sigma}^{g\to h}$, $\breve{\ell}^{g\to h}$, and the 2 time-invariant 2D B-splines GPs each have squared exponential kernels with hyper-parameters $\sigma_{0,i}^{g\to h}$, $\ell_{0,i}^{g\to h}$ and $\ell_{0,j}^{g\to h}$ decomposed as follows,

$$\text{(8)} \qquad k_0^{g\to h}\Big((i,j),(i',j')\Big) = (\sigma_0^{g\to h})^2 \exp\left(-\frac{(i-i')^2}{2(\ell_{0,i}^{g\to h})^2}\right) \exp\left(-\frac{(j-j')^2}{2(\ell_{0,j}^{g\to h})^2}\right).$$

We constrain the model further with a pseudo-likelihood term so that the model's implied incidence rate $\kappa_{r,j}^h$ in (3b) is around the MLE incidence rate estimate obtained from the incidence cohort. We took this approach in lieu of fitting the model to both the source-recipient and individual-level incidence exposure data to bypass extreme computational runtimes, and in the context that the source-recipient data are not informative of incidence dynamics. Specifically, we fitted log-normal distributions to the $1,000 \times 50$ Monte Carlo replicate rate estimates for individuals of gender h and age j in round r (see above) using the `lognorm` R package, and then set

$$\text{(9)} \qquad \frac{\sum_i \lambda_{r,i,j}^{g\to h}}{S_{r,j}^h \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right|} \sim \text{LogNormal}\Big(\text{mean} - \hat{\kappa}_{r,j}^h, \text{var} - \hat{\kappa}_{r,j}^h\Big),$$

where mean-$\hat{\kappa}_{r,j}^h$ and var-$\hat{\kappa}_{r,j}^h$ denote respectively the parameters of the fitted log-normal distributions, and the left-hand side is calculated from (6b) and matches the model's incidence rate $\kappa_{r,j}^h$ in (3b). Model (6-9) was fitted with `Rstan` version

2.21.0, using Stan's adaptive HMC sampler [1]. The applied results of our analyses are available in the preprint [4].

References

[1] Carpenter, B. et al.,*Stan: A probabilistic programming language. Journal of Statistical Software* **76**, 1–32 (2017).

[2] Monod, M. et al., *Regularised B-splines projected Gaussian Process priors to estimate time-trends in age-specific COVID-19 deaths. Bayesian Analysis* **1**, 1–31 (2022).

[3] Xi, X.et al. *Inferring the sources of HIV infection in Africa from deep-sequence data with semi-parametric Bayesian Poisson flow models. Journal of the Royal Statistical Society: Series C (Applied Statistics)* **71**, 517–540 (2022).

[4] Monod, M. et al., *Growing gender disparity in HIV infection in Africa: sources and policy implications. medrxiv* https://doi.org/10.1101/2023.03.16.23287351 (2023).

How immune dynamics shape multi-season epidemics: a continuous-discrete model in one dimensional antigenic space

Mick Roberts

(joint work with Roslyn Hickson, James McCaw)

We extend a previously published model for the dynamics of a single strain of an influenza-like infection [1]. The model incorporated a waning acquired immunity to infection and punctuated antigenic drift of the virus, employing a set of coupled integral equations within a season and a discrete map between seasons. For the within season model we used the Kermack McKendrick equations, those infected entering the removed compartment (R). After spending one season in R, hosts enter a partially susceptible compartment, with probability of being infected k times that for the fully susceptible compartment under the same infection pressure, $k < 1$. To approximate the effects of antigenic drift and population turnover, the proportion of the population in the partially susceptible and removed compartments was multiplied by a constant $c < 1$, and a proportion $1 - c$ was added to the fully susceptible compartment.

The results in [1] show complicated dynamics for a range of parameter values. However, the model does not differentiate between two paradigms: where immunity to infection depends on the time since a host was last infected, and immunity depending on the number of times that a host has been infected. To address this we subdivide the population into a proportion fully susceptible at time t, $S^{\emptyset}(t)$, those partially susceptible $S^1(t) \dots S^{m-1}(t)$ and those removed $S^m(t)$, for some $m > 2$. Those in the S^ℓ compartment are $k_\ell < 1$ as susceptible as those in $S^{\emptyset}$, with $k_m = 0$. At the beginning of a season ($t = 0$), an epidemic *takes off* if

$$\mathcal{R} = \mathcal{R}_0 \left(S^{\emptyset}(0) + \sum_{\ell=1}^{m} k_\ell S^\ell(0) \right) > 1$$

where $\mathcal{R}_0$ is the basic reproduction number. The final size of the epidemic (proportion of the population infected) solves

$$\mathcal{P} = S^{\emptyset}(0)\left(1 - e^{-\mathcal{R}_0\mathcal{P}}\right) + \sum_{\ell=1}^{m} S^{\ell}(0)\left(e^{-\mathcal{R}_0\mathcal{P}} - e^{-\mathcal{R}_0 k_\ell \mathcal{P}}\right)$$

Now define $\mathbf{s}_n$ to be the vector whose ℓ^{th} component is the initial value $S^{\ell}(0)$ in the n^{th} season for $\ell = 1 \ldots m$, and $E(\mathbf{s}_n) = \exp(-\mathcal{R}_0\mathcal{P}_n)$ where $\mathcal{P}_n$ is the final size of the within season epidemic with initial conditions $\mathbf{s}_n$. The between season map becomes

$$\mathbf{s}_{n+1} = \mathbf{C}(E(\mathbf{s}_n))\mathbf{s}_n + \mathbf{q}(E(\mathbf{s}_n))$$

where $\mathbf{C}$ is an $m \times m$ matrix, and $\mathbf{q}$ is an m dimensional vector valued function. We iterated the map for $m = 4$ with $\mathcal{R}_0 = 2.0$, $c = 0.9$ and $k_1 \in (0, 1)$ for the two paradigms. For each example we took $k_2 = k_1^2$ and $k_3 = k_1^3$, with initial condition in the first season $\mathbf{s}_0 = \mathbf{0}$ (entire population susceptible). If immunity depends on the time since last infection, then

$$\mathbf{s}_{n+1} = c \begin{pmatrix} 0 & E^{k_2} & 0 & 0 \\ 0 & 0 & E^{k_3} & 0 \\ 0 & 0 & 0 & 1 \\ E - E^{k_1} & E - E^{k_2} & E - E^{k_3} & E - 1 \end{pmatrix} \mathbf{s}_n + \begin{pmatrix} 0 \\ 0 \\ 0 \\ c\,(1 - E) \end{pmatrix}$$

If immunity depends on the number of infections experienced, then

$$\mathbf{s}_{n+1} = c \begin{pmatrix} E - 1 + E^{k_1} & E - 1 & E - 1 & E - 1 \\ 1 - E^{k_1} & E^{k_2} & 0 & 0 \\ 0 & 1 - E^{k_2} & E^{k_3} & 0 \\ 0 & 0 & 1 - E^{k_3} & 1 \end{pmatrix} \mathbf{s}_n + \begin{pmatrix} c\,(1 - E) \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The results are shown in Figure 1, where we plot the effective reproduction number at the start of each season ($\mathcal{R}$) against k_1.

The results for the situation where immunity depends on the time since infection (Figure 1A) show complicated dynamics for $0.06 \leqslant k_1 \leqslant 0.39$, solutions tend to a fixed point (epidemic of same size every year) for $0.39 < k_1 < 0.63$, and a period two solution (epidemic in alternate years) for $k_1 > 0.63$. For $k_1 < 0.06$ there are four years without an epidemic ($\mathcal{R} < 1$) with an epidemic in the fifth year. Further simulations with different values of $\mathcal{R}_0$ and c (not shown) resulted in similar patterns for most (but not all) parameter combinations. In addition, a simulation with $\mathcal{R}_0 = 2.0$ and $c = 0.9$, but with different initial conditions (also not shown) revealed a second attractor for some values of k_1.

The results for the situation where immunity depends on the number of infections (Figure 1B) show solutions tending to a fixed point (epidemic of the same size every year) for all values of k_1. A similar outcome was observed for all values of $\mathcal{R}_0$ and c investigated. In all cases the fixed point of the map solves

$$\mathbf{s}^{\star} = \left(\mathbf{I} - \mathbf{C}(E^{\star})\right)^{-1} \mathbf{q}(E^{\star})$$

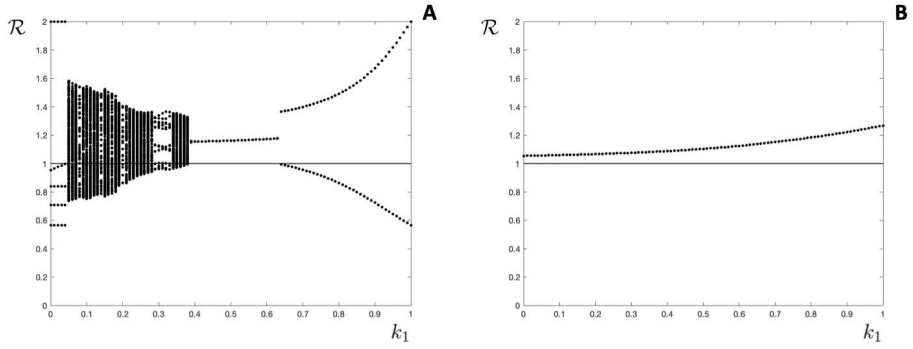

FIGURE 1. Orbit diagrams showing the effective reproduction number $\mathcal{R}$ as a function of the immunity coefficient k_1. A: The time since infection determines immunity. B: The number of infections determines immunity. The map was iterated 1500 times, the last 500 iterations are plotted. The horizontal line is at $\mathcal{R} = 1$.

where $E^\star = \exp(-\mathcal{R}_0 \mathcal{P}^\star)$ and corresponds to an epidemic with final size

$$\mathcal{P}^\star = 1 - E^\star + \sum_{\ell=1}^{m} s_\ell^\star \left(E^\star - E^{\star k_\ell}\right)$$

In summary, if immunity to infection depends on the time since a host was last infected the model exhibited chaotic dynamics in some regions of parameter space, and regions of parameter space with more than one attractor. If immunity depends on the number of times that a host has been infected, the attractor was a stable fixed point corresponding to an identical epidemic each season. We also examined the model with both paradigms in combination, almost but not exclusively observing a stable fixed point or periodic solution. Adding stochastic perturbations to the between season map failed to destroy the model's qualitative dynamics. Our results suggest that if the level of host immunity depends on the elapsed time since the last infection then the epidemiological dynamics may be unpredictable.

REFERENCES

[1] MG Roberts, RI Hickson, JM McCaw, L Talarmain, *A simple influenza model with complicated dynamics*, Journal of Mathematical Biology **78** (2019), 607–624.

Revealing disease ecology from historical records over the last seven centuries

DAVID J. D. EARN

Historical records allow us to reconstruct patterns of disease spread in the past, in some cases going back hundreds of years. The questions we can address depend on the available data, which has varied enormously over time. I presented data, going

back as far as 1348, which we have acquired and studied at McMaster University in the last few years. I discussed the strengths and limitations of the various types of data for mechanistic modelling, and how these data have so far contributed to improving our understanding of infectious disease epidemics.

The largest part of the talk focussed on plague (which is caused by the bacterium *Yersinia pestis*), and more specifically on plague epidemics in London, England. I discussed three types of data from London that we have digitized for analysis over the last 20 years:

- Weekly mortality attributed to plague, as listed in the London Bills of Mortality (LBoM) from 1563 to 1666;
- Weekly mortality from all causes, aggregated from surviving London parish death registers going back to 1538; and
- Daily counts of Last Wills and Testaments written during the 14th century (and probated in the Court of Husting).

I explained how we have used these data to estimate the initial growth rates of plague epidemics in London between 1348 and 1665, and highlighted our recent discovery that plague epidemics in the 17th century grew approximately four times faster than those in the 14th century [3]. I also showed a number of animations of the Great Plague of London in 1665, displaying the initial pattern of spread in early 1665 (from the outskirts to the city centre over several months) and the raging epidemic that ensued by mid summer.

I then described our analysis of the four 19th century cholera pandemics in London [8]. We found that three of the four pandemics (in 1832, 1849, and 1854) were preceded by an out-of-season "herald wave", whereas no such out-of-season wave preceded the 1866 pandemic.

My attention next turned to our analysis of the 1918 influenza pandemic in London. A puzzling feature of that pandemic was the occurrence of three distinct waves within a year. I described our mechanistic modelling of the three waves [4, 5], from which we inferred that the primary mechanism that contributed to generating three waves was likely behavioural change (reducing contact with others) in response to high disease prevalence or mortality. We found that weather and school closures also had detectable effects, but without behavioural response our model was unable to reproduce all three waves.

Finally, I briefly discussed our detailed analysis of weekly smallpox mortality in London between 1664 and 1930 [7]. Over the decades and centuries there were striking changes in smallpox dynamics that were correlated with demographic changes, historical events, and uptake of control measures. We hope to make sense of all of these dynamical changes using mechanistic modelling and analysis tools that have previously allowed us to understand transitions in childhood disease dynamics in the 20th century [2, 1, 6].

I mentioned COVID-19 only in passing in this talk, but all of the types of modelling and analysis that we have applied to the diseases I did discuss are relevant to COVID-19. Some of these techniques have been applied to COVID-19

already, and others—which concern dynamical changes over long timescales—are certain to be applied in the future.

References

[1] C.T. Bauch and D.J.D. Earn, *Transients and attractors in epidemics*, Proceedings of the Royal Society of London, Series B **270** (2003), 1573–1578.

[2] D.J.D. Earn, P. Rohani, B.M. Bolker and B.T. Grenfell, *A simple model for complex dynamical transitions in epidemics*,Science **287** (2000), 667–670.

[3] D.J.D. Earn, J. Ma, H. Poinar, J. Dushoff and B.M. Bolker, *Acceleration of plague outbreaks in the second pandemic*, PNAS – Proceedings of the National Academy of Sciences of the U.S.A. **117(44)** (2020), 27703–27711.

[4] D. He, J. Dushoff, T. Day, J. Ma and D.J.D. Earn, *Mechanistic modelling of the three waves of the 1918 influenza pandemic*,Theoretical Ecology. **4(2)** (2011), 283–288.

[5] D. He, J. Dushoff, T. Day, J. Ma and D.J.D. Earn, *Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales*,Proceedings of the Royal Society of London, Series B. **280(1766)** (2013), 20131345.

[6] K. Hempel and D.J.D. Earn, *A Century of Transitions in New York City's Measles Dynamics*, ournal of the Royal Society of London, Interface **12(106)** (2015), 20150024.

[7] O. Krylova, and D.J.D. Earn, *Patterns of smallpox mortality in London, England, over three centuries*, PLoS Biology **18(12)** (2020).

[8] J.H. Tien, H.N. Poinar, D.N. Fisman and D.J.D. Earn, *Herald waves of cholera in nineteenth century London*, Journal of the Royal Society of London, Interface **58(8)** (2011), 756–760.

Digital Contact Tracing for COVID-19: from Initial Theoretical Evidence to Evaluation

Chris Wymant

(joint work with Luca Ferretti, Michelle Kendall, Daphne Tsallis, Marcos Charalambides, Robert Hinch, Luke Milsom, Matthew Ayres, Lele Zhao, Anel Nurtay, Michael Parker, Chris Holmes, Mark Briers, Lucie Abeler-Dörner, David Bonsall, Christophe Fraser)

In March 2020 we published a proposal to enhance COVID-19 contact tracing by making it digital: using proximity-detecting mobile phone apps to record close contact events [1]. Key motivations were greater speed, automatic scaling of the tracing process with the epidemic size, and detection and memory of contacts unknown to or forgotten by the index case. Our initial mathematical modelling (adapting the double-integral renewal-equation approach of reference [2]) and that of follow-up studies (including our bespoke agent-based modelling) provided the evidence for many countries to develop national digital tracing programmes.

The highly privacy-preserving framework Google and Apple built into their mobile device operating systems to facilitate the intervention securely made epidemiological evaluation difficult. However, with anonymous analytics data collected by the NHS COVID-19 app in England and Wales, we were able to estimate the app's initial epidemiological impact from its launch on 24 September 2020 to the end of December 2020 [3]. It was used regularly by approximately 16.5 million users (28% of the total population), and sent approximately 1.7 million exposure notifications: 4.2 per index case consenting to contact tracing. We estimated that

the fraction of individuals notified by the app who subsequently showed symptoms and tested positive (the secondary attack rate, SAR) was 6%, similar to the SAR for manually traced close contacts. We estimated the number of cases averted by the app using two complementary approaches. First, we performed a regression comparing different spatial areas (lower-tier local authorities), subsetting and matching for comparability, inspired by causal inference methods on big observational data [4]. This method gave an estimate of 594,000 cases averted (95% confidence interval 317,000–914,000). Second, we used mathematical modelling linking the number of individuals traced, the SAR, the proportion of the infectious period reachable by tracing, the effect on the period of quarantining, and the size of downstream transmission chains averted by averting one case at their start. This method gave an estimate of 284,000 cases averted (central 95% range of sensitivity analyses 108,000–450,000). Approximately one case was averted for each case consenting to notification of their contacts. We estimated that for every percentage point increase in app uptake, the number of cases could be reduced by 0.8% (using modelling) or 2.3% (using regression).

References

[1] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing*, Science, **368** (2020).

[2] C. Fraser, S. Riley, R.M. Anderson, N.M. Ferguson, *Factors that make an infectious disease outbreak controllable* PNAS, **101** (2004), 6146–6151.

[3] C. Wymant, L. Ferretti, D. Tsallis, M. Charalambides, L. Abeler-Dörner, D. Bonsall, R. Hinch, M. Kendall, L. Milsom, M. Ayres, C. Holmes, M. Briers, C. Fraser, *The epidemiological impact of the NHS COVID-19 app*, Nature, **594** (2021), 408–412.

[4] M.A. Hernan, J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available*, American Journal of Epidemiology, **183** (2016) 758–764.

The NHS COVID-19 contact tracing app for England and Wales: epidemiological impacts and insights

Michelle Kendall

(joint work with Luca Ferretti, Daphne Tsallis, Andrea Di Francia, Yakubu Balogun, Xavier Didelot, Christophe Fraser)

The NHS COVID-19 app was launched on 24 September 2020 across England and Wales, with millions of users installing it in the first few days after its launch [1]. Its development was motivated by the theoretical finding that rapid, scalable and anonymised contact tracing could help reduce transmission of SARS-CoV-2 [2, 3, 4, 5]. It uses Google and Apple's Bluetooth exposure notification platform [6] to quickly perform contact tracing with the aim of reducing transmission of SARS-CoV-2.

We show that the NHS COVID-19 app's uptake, user engagement, and impact varied according to changing social and epidemic characteristics. We describe

the interaction and complementarity of manual and digital contact tracing approaches. Using anonymised, aggregated app data we show that app users who were recently notified were more likely to test positive than app users who were not recently notified, by a factor that varied considerably over time. We explain how we adapted the modelling approach of Wymant and Ferretti [1] for estimating cases, hospitalisations and deaths averted, building upon the approach to incorporate the background of changing epidemic dynamics including emerging viral variants, population-level restrictions and vaccination roll-out. We estimate that the app's contact tracing function in the first year alone averted about 1 million cases (sensitivity analysis 450,000–1,400,000), corresponding to 44,000 hospital cases (20,000–60,000) and 9,600 deaths (4,600–13,000). These results were recently published [7].

In the second part of the talk we describe results which will be presented in a manuscript which is currently in preparation; these are not to be shared outside of the group of registered participants of the workshop. Here we provide a brief overview.

Although the app is privacy-preserving by design and collects only minimal data to ensure it is functioning correctly, we show that app data has provided valuable insights into the progression of the epidemic, including local and national measures of population contact rates, and the infectiousness of those contact events. Together these have provided a way to directly measure aspects of the reproduction number $R(t)$ in a timely manner, available earlier than estimates of $R(t)$ derived from case numbers or surveys [8]. The decomposition of $R(t)$ into contact rates and infectiousness of contact events enables a more detailed analysis of the drivers of changes in $R(t)$ than is available from case data alone. Finally, we present particular insights from app data concerning regional variations in epidemic dynamics and the impact of behavioural changes during significant national events such as Christmas holidays and major football matches.

References

[1] C. Wymant, L. Ferretti, D. Tsallis, M. Charalambides, L. Abeler-Dörner, D. Bonsall, R. Hinch, M. Kendall, L. Milsom, M. Ayres, C. Holmes, M. Briers, C. Fraser, *The epidemiological impact of the NHS COVID-19 app*, Nature, **594** (2021), 408–412.

[2] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing*, Science, **368** (2020).

[3] M. Kretzschmar, G. Rozhnova, MCJ. Bootsma, M. van Boven, JHHM. van de Wijgert, MJM. Bonten. *Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study*, The Lancet Public Health, 5(**8**):e452-e459 (2020).

[4] AJ. Kucharski, P. Klepac, AJK. Conlan, SM. Kissler, ML. Tang, H. Fry, JR. Gog, WJ. Edmunds, *Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study.*, The Lancet Infectious Diseases, 20(**10**):1151-1160 (2020).

[5] K. Jenniskens, MCJ. Bootsma, JAAG. Damen, M. Ghannad, MS. Oerbekke, RWM. Vernooij, R. Spijker, KGM. Moons, MEE. Kretzschmar, L. Hooft *Effectiveness of contact tracing apps for SARS-CoV-2: an updated systematic review [version 1; peer review: awaiting peer review]*, F1000Research, 11:515 (2022).

[6] Apple and Google. *Apple Google Exposure Notifications: Using technology to help public health authorities fight COVID-19.* (2020) Available from: `https://www.google.com/covid19/exposurenotifications/`
[7] M. Kendall, D. Tsallis, C. Wymant, A. Di Francia, Y. Balogun, X. Didelot, L. Ferretti, C. Fraser, *Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year*, Nature Communications, **14**, 858 (2023).
[8] M. Kendall, C. Fraser *Evaluating epidemiological impacts of the NHS COVID-19 app: an August 2021 update.* Blog post available from: `https://www.coronavirus-fraser-group.org/blog#2august2021`

Parameter estimation from contact-tracing data in graph-based models

Johannes Müller
(joint work with Mirjam Kretzschmar, Augustine Okolie)

We adopt a maximum-likelihood framework based on a stochastic susceptible-infected-recovered (SIR) model with contact tracing on a contact graph that consists of a rooted random tree, as developed in [1]. Given a randomly chosen index case, we are able to find formulas for the distributions of time since infection of that index case, and the number of cases detected by contact tracing, respectively. Based on these results, we derive a maximum likelihood estimator for the parameters of the model. This estimator is influenced by the downstream degree distribution of the underlying contact tree, the basic reproduction number, and the tracing probability.

We inspect different random graph based contact models, that all imply their own degree distributions. As the selected random graphs look locally like trees in case of a large number of nodes, we use our tree-based methods to find those random contact graphs that are consistent with contact tracing data. Thereto, we use a data set from India [2], where simply the number of detectees identified by forward tracing are given. It turns out, that scale free networks, having a power law as degree distribution, is the best explanation of the given data. We not only are able to distinguish between different degree distributions/contact graph structures, but it is also possible to identify the tracing probability. In practical applications, we expect that the estimation of the tracing probability could be a valuable tool in monitoring the efficiency of tracing programs.

References

[1] A. Okolie, J. Müller, *Exact and approximate formulas for contact tracing on random trees*, Math. Biosc. **321** (2020), 108320.
[2] Gupta, M. et al. *Contact tracing of COVID-19 in Karnataka, India: Superspreading and determinants of infectiousness and symptomatic infection*, PLoS ONE **17** (2022), e0270789.

An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters

Edward L. Ionides
(joint work with Ning Ning, Jesse Wheeler, Kidus Asfaw, Jifan Li, Joonha Park and Aaron A. King)

We extend the classes of models for which it is numerically tractable to carry out likelihood-based statistical inference on a collection of partially observed, stochastic, interacting, nonlinear dynamic processes. Each process is called a unit, and our primary motivation arises in biological metapopulation systems where a unit is a spatially distinct sub-population. In this context, the collection of units is called a metapopulation. We consider partially observed Markov processes having this spatiotemporal unit structure, and these are called SpatPOMP models. Likelihood evaluation for SpatPOMP models can be carried out recursively via filtering algorithms. Each increment of a filtering recursion can be broken down into two components, a one-step forecast (in which the latent state propagates according to the Markovian model) and an assimilation step in which a new data point is accounted for by adjusting the forecast. In high dimensional situations (i.e., a more than a few units) the reweighting procedure arising in the assimilation step for the widely-used particle filter fails [2], and so algorithms with improved scalability are required.

Advanced algorithms applied to complex dynamic models require consideration of software implementations, and we discuss these in the context of the R package spatPomp [1]. A range of spatiotemporal filters are coded in spatPomp, and the filtering approach that has proved most effective for metapopulation models is a block particle filter [7]. The block particle filter replaces the The empirical success of this filter on epidemiological metapopulation models was presented by Ionides et al. [3]. Filtering is not sufficient for inference, prima facie, since it provides an evaluation of the likelihood rather than a parameter estimate. However, iterating a filter while perturbing parameters can approach the maximum likelihood estimate (MLE) for a general class of POMP models [4].

We present an iterated block particle filter which provably approximates the MLE for a class of SpatPOMP models while avoiding the "curse of dimensionality" that affects previous iterated filtering algorithms [6]. We demonstrate an extension of this algorithm that is applicable to models having some "shared" parameters (common to all units) and some "unit-specific" parameters which take a distinct value for each unit [5]. We apply this method to study pre-vaccination measles cases in collections of towns in England and Wales, via Susceptible-Exposed-Infectious-Recovered dynamics for each unit (i.e., town) with a gravity model for coupling. We find that unit-specific parameters are required to explain the data, and that gravity coupling does not provide substantially better explanation of these data than an uncoupled model. More importantly, we have demonstrated methods enabling continuation of this investigation in the search for better metapopulation models.

References

[1] K. Asfaw, J. Park, J., A.A. King, and E.L. Ionides, *Statistical inference for spatiotemporal partially observed Markov processes via the R package spatpomp*. arXiv:2101.01157v3. (2023)

[2] T. Bengtsson, P. Bickel and B. Li, *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* (2008), pages 316–334. Institute of Mathematical Statistics, Beachwood, OH.

[3] E.L. Ionides, K. Asfaw, J. Park, and A.A. King, *Bagged filters for partially observed interacting systems*, Journal of the American Statistical Association (2021), pre-published online.

[4] E.L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A.A. King, *Inference for dynamic and latent variable models via iterated, perturbed Bayes maps*, Proceedings of the National Academy of Sciences of the USA, **112(3)** (2015), 719–724.

[5] E.L. Ionides, N. Ning, and J. Wheeler, *An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters*, Statistica Sinica (2022), pre-published online.

[6] N. Ning and E.L. Ionides, *Iterated block particle filter for high-dimensional parameter learning: Beating the curse of dimensionality*, *arXiv:2110.10745* (2021).

[7] P. Rebeschini, P. and R. van Handel, *Can local particle filters beat the curse of dimensionality?*, The Annals of Applied Probability, **25(5)** (2015), 2809–2866.

Statistical and mathematical details of trials for estimating vaccine effectiveness for emerging infectious disease threats

Ira Longini

(joint work with M. Elizabeth Halloran, Claudio Struchiner)

We are facing a global assault of emerging infectious disease threats. Some recent examples are Ebola, Zika, Covid-19, Lassa fever and monkeypox[1, 2]. When these threats emerge, there is an urgent need to evaluate the effectiveness of candidate vaccines as they are rolled out; first as experimental products with unknown effectiveness, and later in terms of optimal deployment for disease control. In this presentation, we develop a series of designs and estimating equations for the evaluation of the direct effectiveness (i.e., efficacy) and the indirect (i.e., herd) protection of these vaccine candidates [3, 4].

We develop a model formulation to estimate the direct, indirect, total, and overall vaccine effects combining data from trials with two types of study designs: individual-randomization within cluster and cluster-randomization, based on a Cox proportional hazards model, where the hazard of infection depends on both vaccine status of the individual as well as the vaccine status of the other individuals in the same cluster [5]. The estimating equations are derived as the partial likelihood score function for the marginal proportional hazards model. Then the estimators for the vaccine effectiveness estimators are derived as functions of the estimated parameters from the proportional hazards model.

We illustrate the use of the proposed model and assess the potential efficiency gain from combining data from multiple trials, compared to using data from each individual trial alone, through two simulation studies, one of which is designed based on a cholera vaccine trial previously carried out in Matlab, Bangladesh [6].

We provide these estimators over a seamless adaptive design for an overall platform vaccine trial [3, 7]. We give further examples of this approach from a past ring vaccine trial for Ebola in Guinea[8], and an upcoming vaccine trial for Lassa fever in Nigeria [9].

References

[1] World Health Organisation (WHO) *An R& D Blueprint for Action to Prevent Epidemics*, https://cdn.who.int/media/docs/default-source/blue-print/an-randd-blueprint-for-action-to-prevent-epidemics.pdf?sfvrsn=f890ab4e_1.

[2] World Health Organisation (WHO) *Prioritizing diseases for research and development in emergency contexts*, https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts.

[3] I.M. Longini, K. Sagatelian, W.N. Rida and M.E. Halloran.*Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations*, Statistics in Medicine 17, 1121–1136 (1998).

[4] M.E. Halloran, I.M. Longini and C.J. Struchiner, *The Design and Analysis of Vaccine Studies*, Springer, New York (2009).

[5] I.M. Longini, Y. Yang, T.R. Fleming, et al. *A platform trial design for preventive vaccines against Marburg virus and other emerging infectious disease threats*, Clinical Trials 19, 647–654 (2022).

[6] I.M Longini IM, A. Nizam, M. Ali, et al. *Controlling endemic cholera with oral vaccines*, Public Library of Science (PloS), Medicine 4 (2007).

[7] N.E. Dean, P. Gsell, R. Brookmeyer, et al. *Creating a framework for conducting randomized clinical trials during disease outbreaks*, New England Journal of Medicine 382, 1366–1369 (2020).

[8] A.M. Henao-Restrepo, I.M. Longini, M. Egger, et al. *Efficacy of a recombinant live VSV-vectored vaccine expressing Ebola surface glycoprotein: Interim results from the Guinea ring vaccination cluster-randomized trial*, The Lancet, 38, 857–866 (2015).

[9] World Health Organisation (WHO) *Efficacy trials of Lassa Vaccines: endpoints, trial design, site selection: WHO Workshop final report (2018)*, https://www.who.int/publications/m/item/efficacy-trials-of-lassa-vaccines-endpoints--trial-design--site-selection--who-workshop.

An epidemic model with short-lived mixing groups

Frank Ball

(joint work with Peter Neal)

Almost all epidemic models make the assumption that infection is driven by the interaction between pairs of individuals, one of whom is infectious and the other of whom is susceptible. However, in society individuals mix in groups of varying sizes, at varying times, allowing one or more infectives to be in close contact with one or more susceptible individuals at a given point in time. In this talk we investigate the effect of mixing groups beyond pairs on the transmission of an infectious disease in an SIR (susceptible $\rightarrow$ infective $\rightarrow$ recovered) model. The talk is based primarily on Ball and Neal [1], which should be consulted for further details.

We consider the following model for the spread of an SIR epidemic, having an infectious period that follows an exponential distribution with mean γ^{-1}, among

a population of size n. Infection is spread via instantaneous mixing events, which occur at the points of a Poisson process having rate $n\lambda$. The sizes of mixing events are independent and identically distributed according to a random variable $C^{(n)}$, which takes values in $\{2, 3, \ldots, n\}$. Suppose that a mixing event has size c. Then c individuals are chosen uniformly at random from the population to form the mixing event. At a mixing event of size c, any infective has probability π_c of making an infectious contact with any given susceptible, with all such contacts occurring independently. Any susceptible that is contacted by at least one infective at a mixing event becomes infected. Infectives cannot infect susceptibles at the mixing event in which they were infected. Initially there are m_n infectives and $n - m_n$ susceptibles. The epidemic ends when there is no infective remaining in the population. The same model was introduced independently by Cortez [2]. If all mixing events have size 2 (i.e. $\mathrm{P}(C^{(n)} = 2) = 1$), the model is identical to the standard homogeneously mixing stochastic SIR epidemic with individual-to-individual infection rate $\frac{2\lambda\pi_2}{n-1}$ and recovery rate γ.

We consider sequences of epidemic processes, indexed by n, under the assumption that $C^{(n)} \xrightarrow{\mathrm{D}} C$ as $n \to \infty$, where $\mathrm{P}(C = c) = p_C(c)$ $(c = 2, 3, \ldots)$. Further assumptions required for the results are given in Ball and Neal [1].

The early stages of an epidemic with few initial infectives can be approximated by a branching process, which assumes every mixing event that contains infectives has one infective with all others at the mixing event being susceptible. The basic reproduction number R_0 for the epidemic is given by the mean of the offspring distribution of this branching process, i.e.

$$R_0 = \frac{\lambda}{\gamma} \sum_{c=2}^{\infty} \pi_c c(c-1) p_C(c).$$

The early exponential growth rate of the epidemic is $r = \gamma(R_0 - 1)$, which coincides with that of the standard homogeneously mixing SIR epidemic. Let z denote the extinction probability of this branching process assuming one initial individual. For sufficiently large n, the probability that an epidemic with one initial infective takes off and leads to a major outbreak is approximately $1 - z$.

Let $S^{(n)}(t)$ and $I^{(n)}(t)$ be the numbers of susceptibles and infectives at time t. Then $\{(S^{(n)}(t), I^{(n)}(t)) : t \geq 0\}$ is an (asymptotic) density dependent population process. Exploiting the theory of such processes yields the following law of large numbers for $\{(S^{(n)}(t), I^{(n)}(t)) : t \geq 0\}$. Suppose that $n^{-1} m_n \to \epsilon$ as $n \to \infty$, where $\epsilon > 0$. Then, for any $t_0 > 0$,

$$\sup_{0 \leq t \leq t_0} \left| n^{-1}(S^{(n)}(t), I^{(n)}(t)) - (x(t), y(t)) \right| \xrightarrow{\mathrm{p}} 0 \quad \text{as} \quad n \to \infty,$$

where $\{(x(t), y(t)) : t \geq 0\}$ satisfies the following ODE

$$(1) \qquad \frac{dx}{dt} = -\lambda x g(y), \qquad \frac{dy}{dt} = \lambda x g(y) - \gamma y, \qquad (x(0), y(0)) = (1 - \epsilon, \epsilon),$$

with

$$g(y) = \sum_{c=2}^{\infty} p_C(c)c\left[1-(1-y\pi_c)^{c-1}\right].$$

A functional central limit theorem for fluctuations of $\{n^{-1}(S^{(n)}(t), I^{(n)}(t)) : t \geq 0\}$ about $\{(x(t), y(t)) : t \geq 0\}$ is given in Ball and Neal [1].

Let $T^{(n)} = n - S^{(n)}$ be the total size of the epidemic. By exploiting a random time-scale transformation of $\{(S^{(n)}(t), I^{(n)}(t)) : t \geq 0\}$ we obtain the following law of large numbers for $T^{(n)}$. (A corresponding central limit theorem is given in Ball and Neal [1].) Suppose that $n^{-1}m_n \to \epsilon$ as $n \to \infty$, where $\epsilon > 0$, and without loss of generality that $\gamma = 1$. Then $n^{-1}T^{(n)} \xrightarrow{\mathrm{p}} \tau_\epsilon$ as $n \to \infty$, where $\tau_\epsilon = \inf\{t > 0 : \tilde{y}(t) = 0\}$ and $\tilde{y}(t)$ is the solution of the ODE

$$\frac{d\tilde{y}}{dt} = \lambda(1-\tilde{y}-t)\tilde{g}(\tilde{y}) - 1, \qquad \tilde{y}(0) = \epsilon,$$

where $\tilde{g}(y) = y^{-1}g(y)$, if $y \neq 0$ and $\tilde{g}(0) = \lambda^{-1}R_0$. If $m \in \mathbb{N}$ is fixed and $m_n = m$ for all $n \geq m$, then $n^{-1}T^{(n)} \mid T^{(n)} > \log n \xrightarrow{\mathrm{p}} \tau_0$ as $n \to \infty$, where τ_0 is obtained by setting $\epsilon = 0$ in the above ODE. (Note that under these conditions, $\mathrm{P}(T^{(n)} > \log n) \to 1 - z^m$ as $m \to \infty$.)

We now assume that $\pi_c = \pi$ for all c and consider epidemics with fixed R_0 and γ. Write z and τ_ϵ as $z(R_0, C, \pi)$ and $\tau_\epsilon(R_0, C, \pi)$, respectively. (In the sequel, it is assumed implicitly that $R_0 > 1$ if $\epsilon = 0$.) When $\mathrm{P}(C = 2) = 1$, the extinction probability z and the final size τ_ϵ are independent of $\pi \in (0, 1]$; denote them by $\hat{z}(R_0)$ and $\hat{\tau}_\epsilon(R_0)$, respectively. These are the extinction probability and final size for the standard SIR model. For fixed R_0 and event size distribution C, the extinction probability $z(R_0, C, \pi)$ increases with π and the final size $\tau_\epsilon(R_0, C, \pi)$ decreases with π. Further, $z(R_0, C, \pi) \downarrow \hat{z}(R_0)$ and $\tau_\epsilon(R_0, C, \pi) \uparrow \hat{\tau}_\epsilon(R_0)$ as $\pi \downarrow 0$.

To compare epidemics with common infection probability π but different event size distributions C and C', recall PGF ordering ($\overset{g}{\leq}$) of random variables:

$$C' \overset{g}{\leq} C \quad \text{if and only if} \quad f_{C'}(s) \geq f_C(s) \quad \text{for all } 0 \leq s \leq 1,$$

where for a random variable C, $f_C(s) = \mathrm{E}[s^C]$ denotes its probability-generating function. For a random variable C taking values in $\{2, 3, \dots\}$, with $\mathrm{E}[C^2] < \infty$, let $\hat{C}$ be the random variable with "size-biased" distribution

$$\mathrm{P}(\hat{C} = c) = \frac{p_C(c)c(c-1)}{\mathrm{E}[C(C-1)]} \qquad (c = 2, 3, \dots).$$

Suppose that $\hat{C}' \overset{g}{\leq} \hat{C}$. Then, $z(R_0, C', \pi) \leq z(R_0, C, \pi)$, with strict inequality if $R_0 > 1$ and $C \overset{D}{\neq} C'$. Further, $\tau_\epsilon(R_0, C', \pi) \geq \tau_\epsilon(R_0, C, \pi)$, with strict inequality if $C \overset{D}{\neq} C'$. Note that if $\mathrm{P}(C' = 2) = 1$ then $C' \overset{g}{\leq} C$ for any random variable C taking values in $\{2, 3, \dots\}$. Thus the extinction probability and final size of the model with mixing events are respectively greater than and less than those of the standard SIR model with the same R_0.

We now consider epidemics in which all mixing events have size c and $\epsilon > 0$ is small. Letting $\pi = \pi_c$, it follows from the second ODE in (1) that $y(t) \leq \frac{R_0}{(c-1)\pi}$ for all $t \geq 0$. If $R_0 > 1$ and π are held fixed, the peak of an epidemic decreases with c and its duration increases. Moreover, if c is large, an epidemic has very long duration and its size is only marginally greater than the herd immunity level $1 - R_0^{-1}$.

If an exposed period having length which follows an exponential distribution with mean δ^{-1} is added to the above model, R_0 and the extinction probability z remain unchanged, though the early exponential growth rate does change, as does the final size τ_ϵ, which decreases with δ and tends to that of the SIR model as $\delta \to \infty$. As $\delta \downarrow 0$, the final size τ_ϵ increases to $\hat{\tau}_\epsilon(R_0)$.

Finally, we briefly consider adding demography to the above SIR model. We assume that individuals are born at rate μn, where n now is the equilibrium population size in the absence of disease, and die at rate μ, independent of disease status. All newborns are assumed to be susceptible. Mixing events now occur at rate $\lambda[S^{(n)}(t) + I^{(n)}(t) + R^{(n)}(t)]$, where $R^{(n)}(t)$ is the number of recovered individuals at time t. For epidemics with fixed R_0, γ and μ, and constant event size c, the probability of fade out after the first wave of infection decreases with c and increases with π, provided $c > 2$. Let $\mathrm{E}[T_Q^{(n)}]$ be the mean time to disease extinction starting from the quasi-stationary distribution of $\{(S^{(n)}(t), I^{(n)}(t), R^{(n)}(t)) : t \geq 0\}$. Using a functional central limit theorem for $\{(S^{(n)}(t), I^{(n)}(t), R^{(n)}(t)) : t \geq 0\}$ to approximate that quasi-stationary distribution, and hence $\mathrm{E}[T_Q^{(n)}]$, for large n, we find that $\mathrm{E}[T_Q^{(n)}]$ decreases with both c and π.

References

[1] F. Ball and P. Neal, *An epidemic model with short-lived mixing groups*, Journal of Mathematical Biology **85**, 63 (2022), https://doi.org/10.1007/s00285-022-01822-3.

[2] R. Cortez, *SIR model with social gatherings*, arXiv:2203.08260 (2022)

Semiparametric Inference of the Effective Reproduction Number Dynamics from Wastewater Gene Counts with Minimal Compartmental Models

Isaac H. Goldstein

(joint work with Daniel Parker, Sunny Jiang, Volodymyr M. Minin)

Introduction

Pathogen RNA counts collected from wastewater have recently become available as a new data source to use when modelling the spread of infectious diseases. There is a need for new statistical models which can use this data source effectively. Compartmental models, such as the SEIRS model, are one possible class of models. Infected individuals can emit pathogen RNA well after the end of the infectious period. Compartmental models can account for this characteristic

through an additional recovered but still emitting compartment in the model (the SEIRRS model). One drawback of this class of models is that assumptions must be made about the initial number of susceptible individuals and the rate at which recovered individuals become susceptible again. These assumptions are often unverifiable and can greatly influence inference of model parameters, including the effective reproduction number. Inspired by birth-death modelling in infectious disease phylodynamics, we propose an alternative model (the EIRR model), where the time-varying immigration rate into the E compartment can be interpreted as a compound parameter equal to the product of the proportion of susceptibles in the population and the transmission rate. This model allows us to correctly estimate the effective reproduction number while avoiding difficult to verify assumptions about the susceptible population. We apply our new model to estimating the effective reproduction number of SARS-CoV-2 in Los Angeles, California, using pathogen RNA collected from a large wastewater treatment facility.

Methods

Pathogen RNA Count Data. Suppose counts of pathogen are collected from wastewater treatment facilities a total of T times. It is common practice to have multiple measurements taken from the same sample of wastewater at the same time. Often, an average taken across the multiple measurements is reported. These measurements are called replicates. We define $\mathbf{X_j} = (X_{t_1,j}, X_{t_2,j}, \ldots, X_{t_T,j})$ where $X_{t_i,j}$ be the jth replicate of the counts of the gene collected at time t_i. We will model $X_{t_i,j}$ as a noisy representation of the unobserved total number of currently infectious and recently recovered individuals.

The EIR Model. The classic SEIR model models a population moving through four stages susceptible (S), infected but not yet infectious (E), infectious (I) and recovered (R). For the EIR model, we define α_t to be the product $\beta_t \times S/N$. The deterministic EIR model is described with an abbreviated system of ordinary differential equations:

$$\begin{aligned}\frac{dE}{dt} &= \alpha_t \times I - \gamma \times E,\\ \frac{dI}{dt} &= \gamma \times E - \nu \times I,\\ \frac{dR}{dt} &= \nu \times I.\end{aligned}$$

The rate of new latent infections is an immigration rate which does not depend on the S compartment. Importantly, using the EIR model, the effective reproduction number is recoverable, as $R_t = \frac{\beta_t S(t)}{\nu N} = \frac{\alpha_t}{\nu}$. The EIR model allows us to avoid having to specify the initial number of susceptibles and rates of recovery.

modelling RNA gene counts collected from wastewater. We follow Nourbakhsh (2022) in modelling SARS-CoV-2 RNA gene counts as a realization of both currently infectious and recently recovered individuals. This requires a revision of our ODE model. We split the R compartment in two, with equations for each compartment:

$$\frac{dR1}{dt} = \nu \times I - \eta \times R1,$$
$$\frac{dR2}{dt} = \eta \times R1.$$

The $R1$ compartment represents individuals who are no longer infectious, but are still shedding pathogen RNA via fecal matter. With this modified ODE, we model the log of pathogen RNA counts as follows:

$$\log X_{t_i,j} \sim \text{Generalized T}(\log(I(t_i) * \lambda + (1-\lambda) * R1(t_i)) + \log(\rho), \tau^2, df).$$

Here $I(t_i)$ is the number of currently infectious individuals at time t_i, $R1(t_i)$ is the number of non-infectious but still shedding individuals at time t_i. The parameter λ represents, at an individual level, the proportion of shedding which occurs during the infectious period. We use the parameter ρ to allow for flexibility in translating between counts of individuals and counts of RNA. Parameter τ accounts for variation from the mean, and df is the parameter governing the degrees of freedom of the T distribution.

modelling the time-varying reproduction number. We use a random walk prior for the effective reproduction number.

$$\begin{aligned} R_0 &\sim \text{Log-Normal}(\mu_0, \sigma_0), \\ \sigma &\sim \text{Log-Normal}(\mu_{rw}, \sigma_{rw}), \\ \log(R_{k_i}) | R_{k_{i-1}}, \sigma &\sim \text{Normal}(\log(R_{k_{i-1}}), \sigma). \end{aligned}$$

We also use priors on the initial compartment sizes, and on all other model parameters.

We use Hamiltonian Monte Carlo, implemented in the `Julia` package `turing` to sample from the target posterior distribution.

Results

The delta variant was introduce in Los Angeles in June 2021, the omicron variant in November 2021, which are reflected in increases in the counts of pathogen RNA seen in the data (**Figure 1**). The posterior estimates of the effective reproduction number show peaks in mid July 2021 and early December 2021, with wide credible intervals at both peaks.

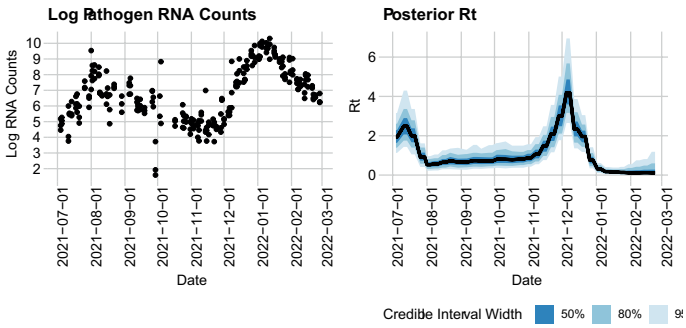

FIGURE 1. Estimation of the effective reproduction number of SARS-CoV-2 in Los Angeles, CA from July 2021 through February 2022. The left plot features the pathogen RNA counts collected from wastewater samples used to create the estimate, displayed in the right plot. Blue regions represent credible intervals, while black lines represent the posterior median.

DISCUSSION

Currently, our method provides quite uncertain estimates of the reproduction number, reflecting high noise levels in the data. Future work will explore how these noise levels might be reduced by more sophisticated priors on the effective reproduction number and through simulations testing how many replicates would be needed to reduce the noise produced during data collection. Currently, we ignore factors related to the sewer system such as changes in population or changes in water flow. Our model could be adapted to allow for covariates to account for these changes.

REFERENCES

[1] S. Nourbakhsh, et al. *A wastewater-based epidemic model for sars-cov-2 with application to three canadian cities*, Epidemics, **39:100560** (2022).

Open science approaches to the mathematical modelling of infectious disease

SIMON FROST

A large number of mathematical and computational models of infectious disease have been developed, particularly in the last fifty years. Despite this rich resource, numerical analysis of these models by third parties is complicated by many factors, including the lack of detail in how the model was developed and the numerical methods used, and the use of unmaintained software packages. Models of infectious disease can be made more findable, accessible, interoperable, and reproducible through a combination of; (a) specific formulations of the models; (b) packaging

of the model and any dependencies; and (c) adding a programming interface to these models.

Construction of a compartmental model of disease transmission usually starts by writing down equations for the rates of change of each compartment, e.g. (S)usceptibles, (I)nfecteds, and (R)ecovereds. However, a more flexible approach starts with writing equations for the processes underlying the rates of change e.g. transmission and recovery, an approach widely taken e.g. in chemical kinetics models. This representation of Markovian models can be easily converted into ordinary differential equations, stochastic differential equations, or jump processes, and enables the automated construction of models for the moments and the probabilities of a given state of the stochastic system at a particular time. Many models are complex, but can be made simpler to understand if they are constructed from smaller building blocks, wired together using a causal approach, where outputs from one submodel are fed into the inputs of another model, or an acausal approach, with potentially bidirectional inputs/outputs, and where common states in the submodels are identified. Use of specific software can also make models easier to use e.g. the use of domain-specific languages to describe a model that make the computer program look more similar to the mathematical equations, or more powerful, e.g. the use of software that can use automatic differentiation to calculate gradients, allowing techniques such as local sensitivity analysis and inference using Hamiltonian Monte Carlo to be used.

For models with numerical results, the written equations are an incomplete description of the entire system. The numerical precision used, the numerical solver used (along with the associated discretization of continuous variables), and even the underlying computer package can all affect the output. In addition, software systems evolve over time, and may become unmaintained. To overcome this, a model along with all the dependencies and data required to run the model should be packaged together in order to allow a third party to exactly replicate the results of a model. One technique to do this is to use a 'container', a kind of lightweight virtual machine. One advantage of this approach is that it also allows running the model in cloud computing environments, allowing much bigger models and larger parameter spaces to be considered.

Even with a model running in a container, a third party is likely to want to be able to change the parameter values of the model and observe changes in the associated output. Each model may have its own formats for model inputs and outputs, making it time consuming as well as potentially error-prone to run a model. One potential remedy is to build an interface to the model through the use of Uniform Resource Identifiers (URIs), which are used to interact with web pages. Not only does this allow running of a model through a web browser or a command line, but it also allows verification of model inputs, reducing the risk that misleading outputs will arise as a consequence of inappropriate choice of input parameters. Modern computer tools allow this to be done with a few extra lines of code.

By following these open science practices, modelers can take advantage of platforms that can run models in an automated way, such that multiple models can be more easily compared. Interoperability of models means that developments e.g. in modelling transmission or in numerical approaches can be quickly and easily incorporated. By defining models in terms of their inputs and outputs, modelers can work with e.g. data scientists to build platforms that feed data to the models, without the data scientists needing to know the details of the model architecture. Through an open science approach, theoretical and applied model development can be accelerated and distributed more widely, increasing its impact in multiple areas of scientific endeavour.

Reporter: Pieter Trapman

Participants

Prof. Dr. Kari Auranen
Department of Mathematics and Statistics
Turku University
20014 University of Turku
FINLAND

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

Dr. Martin Bootsma
Budapestlaan 6
Department of Mathematics
Utrecht University
P.O. Box 80.010
3508 TA Utrecht
NETHERLANDS

Prof. Dr. Tom Britton
Department of Mathematics
Stockholm University
10691 Stockholm
SWEDEN

Andrea Brizzi
Imperial College London
Department of Mathematics
Huxley Building
180 Queen's Gate
London SW7 2AZ
UNITED KINGDOM

Prof. Dr. Caroline Colijn
Dept. of Mathematics and Statistics
Simon Fraser University
Burnaby BC V5A 1S6
CANADA

Prof. Dr. Ben Cooper
Centre for Tropical Medicine and Global Health, New Richards Building, Old Road Campus
Oxford University
Roosevelt Drive
Oxford OX3 7LG
UNITED KINGDOM

Prof. Dr. David Earn
Dept. of Mathematics & Statistics
McMaster University
1280 Main Street West
Hamilton ON L8S 4K1
CANADA

Prof. Dr. John Edmunds
London School of Hygiene and Tropical Medicine
University of London
Keppel Street
London WC1E 7HT
UNITED KINGDOM

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

Prof. Dr. Martin Eichner
Institut für Klinische Epidemiologie und Angewandte Biometrie
Universität Tübingen
Silcherstraße 5
72070 Tübingen
GERMANY

Prof. Dr. Simon Frost
Microsoft Research
Microsoft Building 99
14820 NE 36th St
Redmond WA 98052
UNITED STATES

Isaac Goldstein
Department of Statistics
University of California, Irvine
708 Verano Place
Irvine, CA 92617
UNITED STATES

Prof. Dr. M. Elizabeth Halloran
Department of Biostatistics
University of Washington and
Fred Hutchinson Research Center
1100 Fairview Ave. N, M2-C200
Seattle, WA 98109-1024
UNITED STATES

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

Prof. Dr. Edward L. 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
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Dr. Eben Kenah
Division of Biostatistics
College of Public Health
The Ohio State University
1841 Neil Avenue
Columbus, OH 43210
UNITED STATES

Dr. Michelle Kendall
Department of Statistics
University of Warwick
Gibbet Hill Road
Coventry CV4 7AL
UNITED KINGDOM

Prof. Dr. Aaron King
Department of Ecology and Evolutionary Biology
3038 Biological Sciences Building
1105 North University Avenue
Ann Arbor MI 48109-1048
UNITED STATES

Dr. Don Klinkenberg
National Institute for Public Health and the Environment
RIVM
P.O. Box 1
3720 BA Bilthoven
NETHERLANDS

Prof. Dr. Mirjam Kretzschmar
Department of Epidemiology,
University Medical Center Utrecht
Heidelberglaan 100
3584CX Utrecht
NETHERLANDS

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

Dr. KaYin Leung
National Institute for Public Health and the Environment
RIVM
P.O.Box 1
3720 BA Bilthoven
NETHERLANDS

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

Prof. Dr. Emma McBryde
Australian Institute of Tropical Health and Medicine
James Cook University
James Cook Drive,
Townsville QLD 4811
AUSTRALIA

Dr. Joel Miller
Department of Mathematics
La Trobe University
Bundoora Victoria 3086
AUSTRALIA

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

Nicola Mulberry
Department of Mathematics
Simon Fraser University
8888 University Dr W
Burnaby V5A 1S
CANADA

Prof. Dr. Johannes Müller
Zentrum Mathematik
Technische Universität München
Boltzmannstraße 3
85748 Garching bei München
GERMANY

Dr. Nicola Müller
Vaccine and Infectious Disease Division
Fred Hutchinson Cancer Center
1100 Fairview Ave N
Seattle WA 98109
UNITED STATES

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

Prof. Dr. Julia Palacios
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

Dr. Lorenzo Pellis
School of Mathematics
The University of Manchester
Alan Turing Building
Oxford Road
Manchester M13 9PL
UNITED KINGDOM

Dr. Oliver Ratmann
Faculty of Natural Sciences
Department of Mathematics and Statistics
Imperial College London
South Kensington Campus
525 Huxley Building
London SW7 2AZ
UNITED KINGDOM

Prof. Dr. Mick G. Roberts
New Zealand Institute for Advanced Study
Massey University
North Shore City Mail Centre
Private Bag 102904
0745 Auckland
NEW ZEALAND

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

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

Dr. Claudio J. Struchiner
Escola de Matemática Aplicada
Fundacao Getúlio Vargas
Rua Bejamin Batista 22/202
Rio de Janeiro 22461-120
BRAZIL

Alice Thompson
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Panayiota Touloupou
School of Mathematics
The University of Birmingham
Edgbaston
Birmingham B15 2TT
UNITED KINGDOM

Dr. Pieter Trapman
Mathematisch Instituut
Rijksuniversiteit Groningen
Postbus 800
9700 AV Groningen
NETHERLANDS

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

Prof. Dr. Jacco Wallinga
Centre for Infectious Disease Epidemiology
National Institute for Public Health and the Environment (RIVM)
P.O. Box 1
3720 BA Bilthoven
NETHERLANDS

Dr. Daniel Wilson
Big Data Institute, Oxford Population Health
Old Road Campus
Oxford OX3 9DU
UNITED KINGDOM

Prof. Dr. Martin Wolkewitz
Institute of Medical Biometry and Statistics
Faculty of Medicine and Medical Center
University of Freiburg
Ernst-Zermelo-Str. 1
79104 Freiburg i. Br.
GERMANY

Dr. Chris Wymant
Big Data Institute
Nuffield Department of Medicine
Li Ka Shing Centre for Health Information
and Discovery
University of Oxford
Oxford OX3 7LF
UNITED KINGDOM

Dr. Jason Xu
Department of Statistical Science
Duke University
214 Old Chemistry
Durham, NC 27708-0251
UNITED STATES

Dongni Zhang
Department of Mathematics
Stockholm University
106 91 Stockholm
SWEDEN