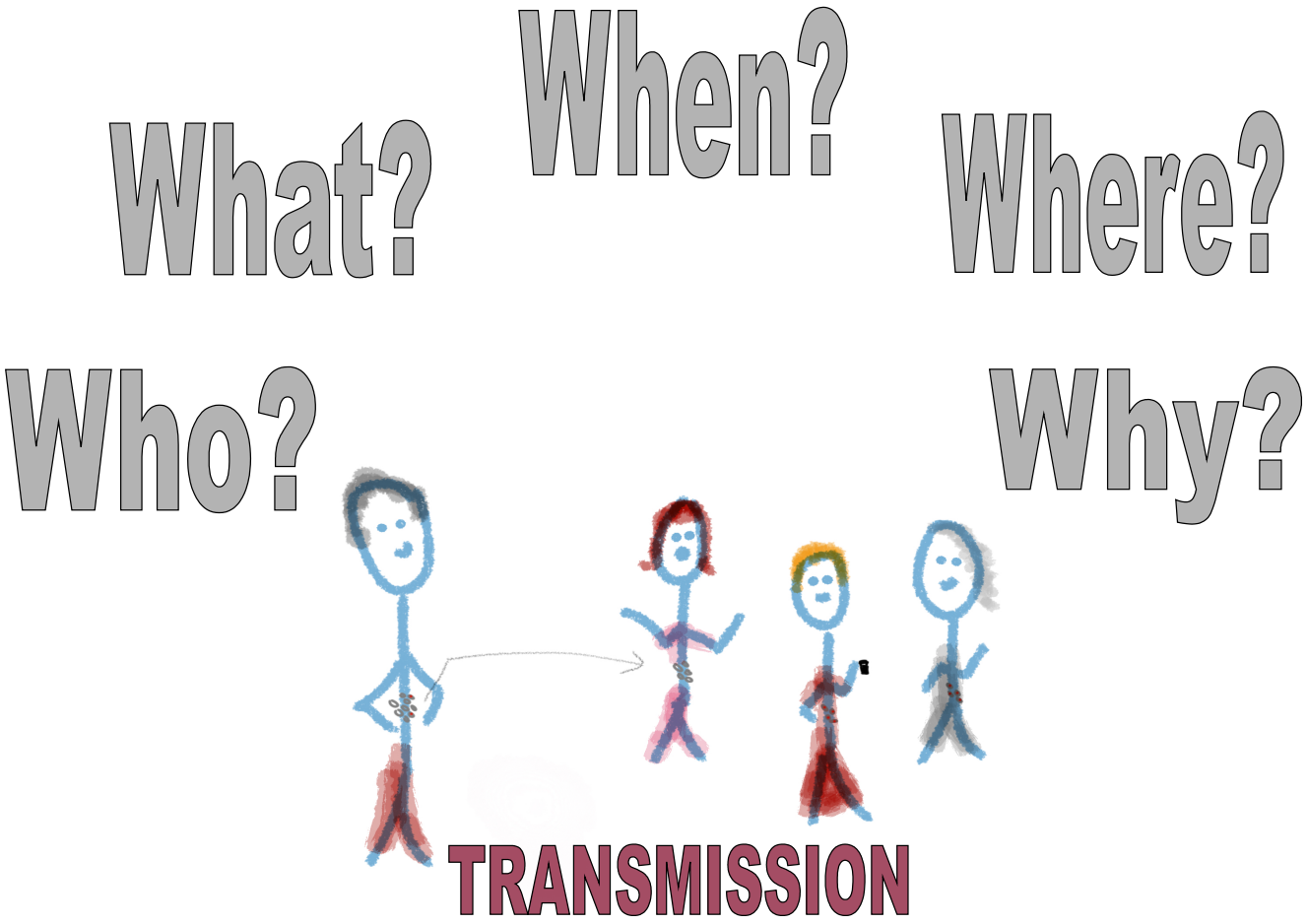
Bayesian Outbreak Reconstruction with Unsampled Cases and Tree Uncertainty

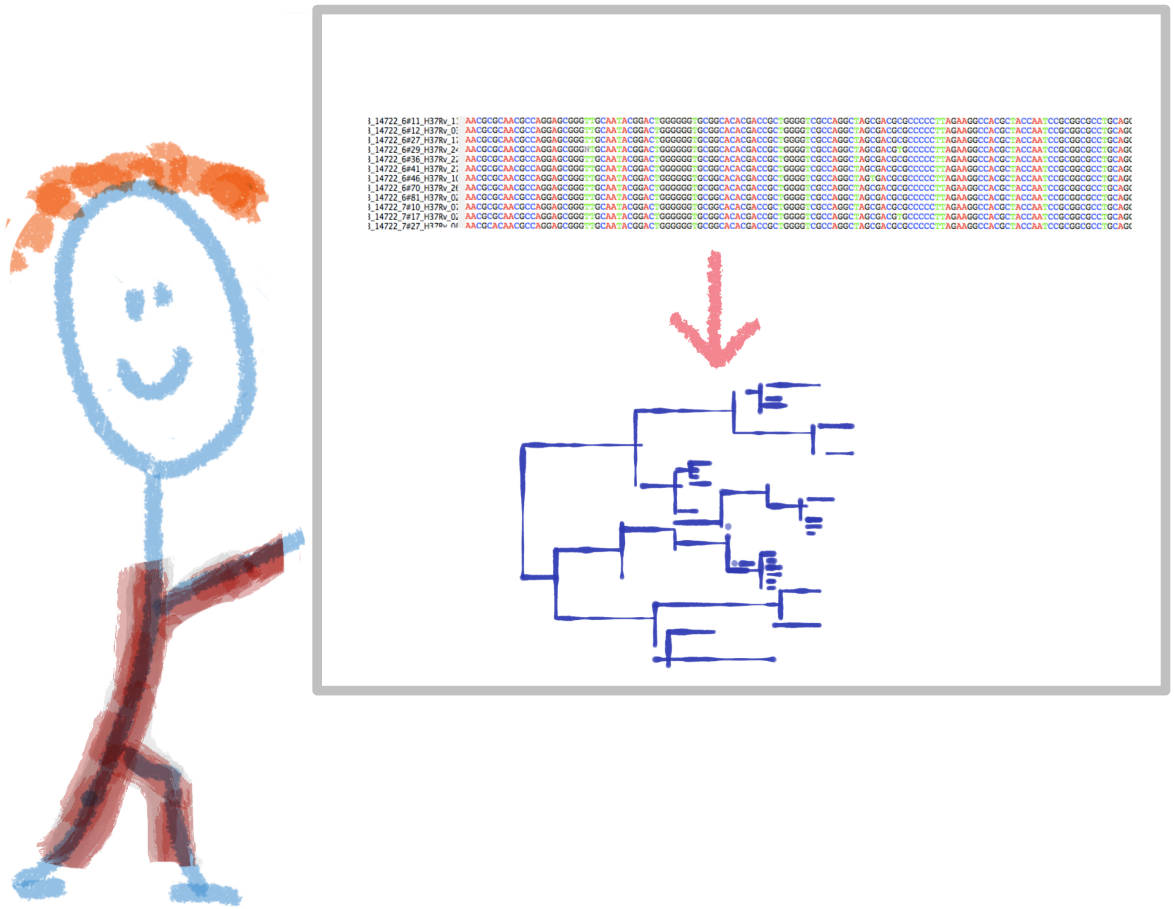
Caroline Colijn

Imperial College London
Department of Mathematics

We want to understand transmission



We have sequences



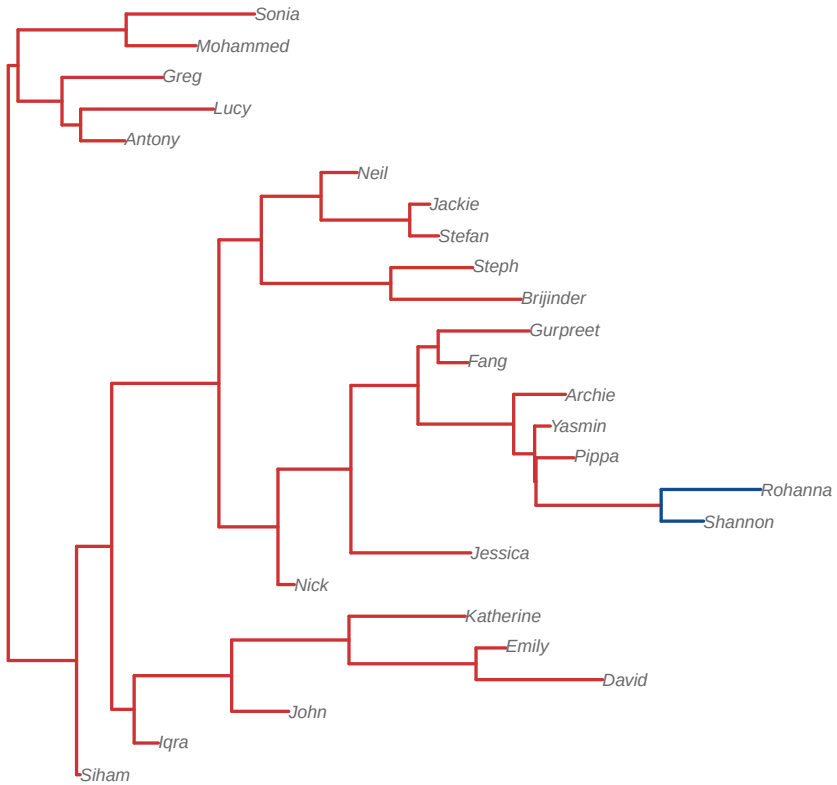
Transmissions cannot all be phylogenetic pairs

You can only be in a pair with one other.

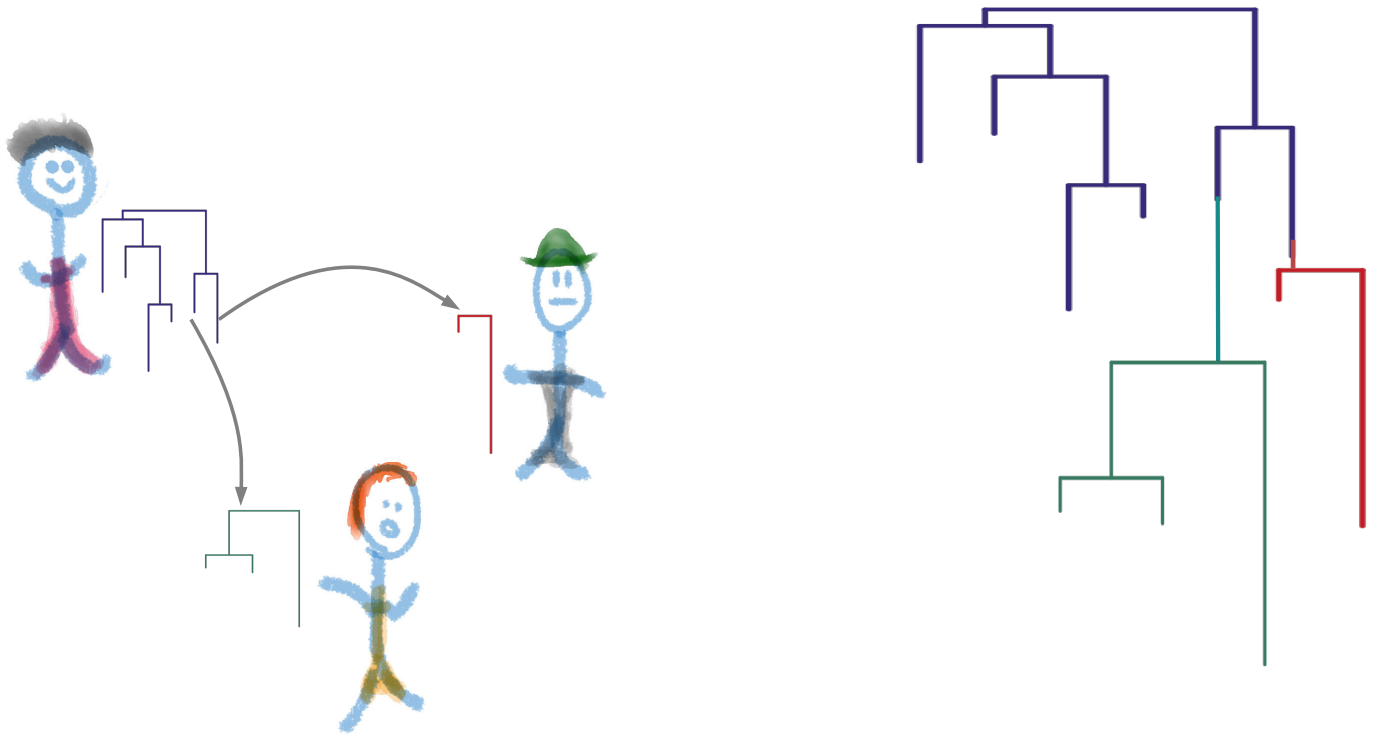
But you could infect many others.

Not all transmission events can be phylo pairs.

Not all closely related pairs of sequences are transmission events.



Genomics doesn't directly reveal transmission



Orange-hair and green-hat are a pair in the phylogeny, but neither infected the other. Both were infected by grey-head.

Bayesian outbreak reconstruction with unsampled cases

TransPhylo: Inference of who infected whom from sequence data

- with in-host diversity
- with unsampled cases
- flexible probability to be sampled
- flexible generation time
- flexible time to sampling

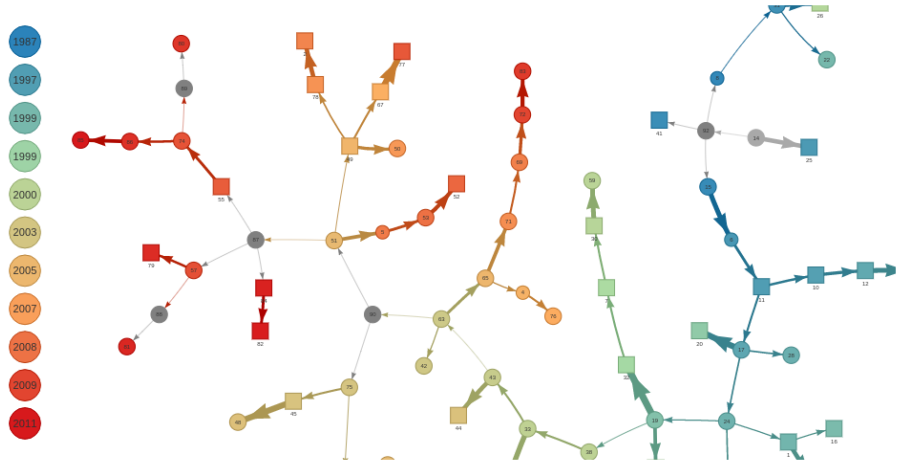
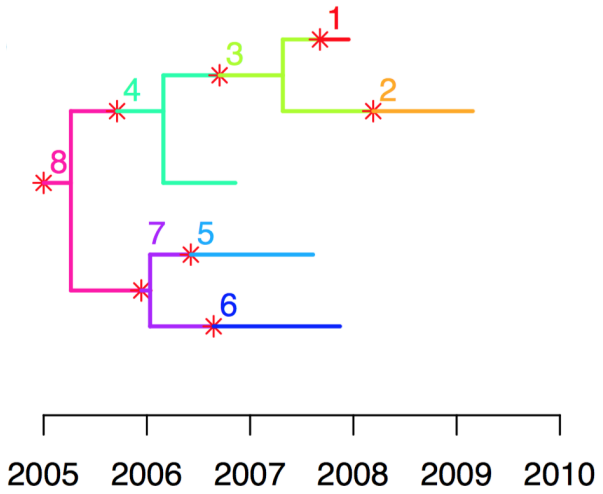


Illustration of who infected whom with uncertainty and unsampled cases

How does it work?

- Hosts can harbour more than one branch of the tree at a time
- Each branch can only be in *one* host at each time
- Branches change hosts at transmission events.
- There can be unsampled hosts.



Colour: which host a lineage is in

TransPhylo is an MCMC method, using augmentation.
It computes the likelihood of the transmission tree (colours)

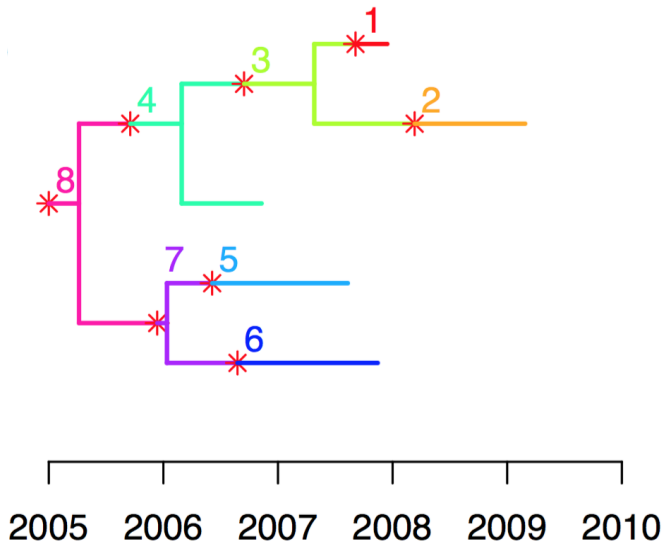
We use the colouring to find a likelihood

TRANSMISSION

- Colour changes are transmission events
- Likelihood: branching process model (for now)

PHYLOGENIES

- Each in-host tree is one colour
- Each in-host tree is *independent* of the others



One colour per host

Decomposition given fixed phylogenetic tree

$$\mathcal{L}(\text{Trans}|\text{Phylo}) \propto \mathcal{L}(\text{Trans events})\mathcal{L}(\text{Phylo}|\text{Trans events})\textit{Priors}$$

$\mathcal{L}$(Transmissions):

- Epidemic model for the system: latency, time to infection, time to sampling
- Finite time due to study end (or the present): this modifies the distribution secondary cases depending on infection time and the sampling probability

$\mathcal{L}$(Phylo|Trans events) :

- Each colour is independent: many little trees
- Coalescent for each one

Math: finite time, unsampled cases

Need: probability unsampled and no sampled descendants

$$p_0(t) = \text{P(unsampled)} \sum_k \text{P(k offspring at } \tau_j\text{) P(they are unsampled)}$$

$$= (1-\pi_t) \sum_{k=0}^{\infty} p_k(t) \prod_{j=1}^{k} \left[\int_t^{\infty} f_g^T(\tau_j - t) p_0(\tau_j) d\tau_j \right] \tag{1}$$

This gives us an **integral equation** for the unknown function $p_0(t)$:

$$p_0(t) = (1-\pi_t) \left(\frac{1-p}{1-p\bar{p}_0(t)} \right)^r .$$

Solved with **the trapezoid method**, exploiting the assumption $f_g(0) = 0$. Then build $\mathcal{L}$(Transn tree) up using $p_0(t)$.

Method sketch

Start with a phylogenetic tree (units of time) and info for your epidemiological model.

1. Propose a colouring: who infected whom, and when
2. Compute its likelihood using the epidemiology model
   - how long between infection and sampling? natural history? R0?
   - handle unsampled cases (challenging)
3. Compute the likelihood for the mini-trees inside each host
4. Accept or reject the proposal
5. Continue (MCMC)

At the end you have a posterior collection of who infected whom and when transmission trees.

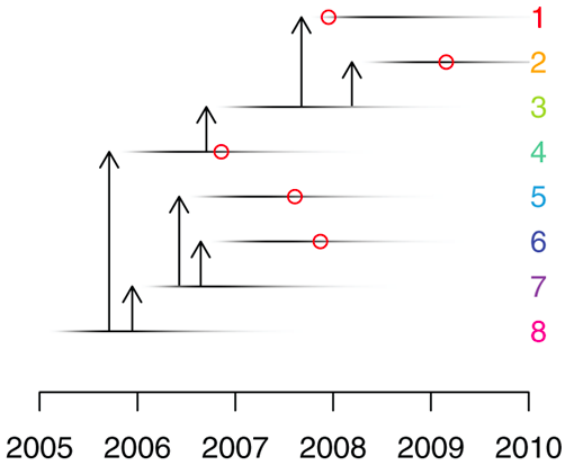
What data does TransPhylo need?

- A timed phylogenetic tree (or a posterior collection of them)
- Sampling dates for the tips (ie the isolates)
- A prior for the time between getting infected and infecting someone else
- A prior for the time between getting infected and getting sampled
- A prior for the overall probability of being sampled eventually
- The time when sampling stopped. Finite time makes a difference! (censoring)

What does TransPhylo produce?

A posterior collection of

- transmission trees: who infected whom?
- with generation times
- with times between infection and sampling
- with unsampled cases and their locations in the phylogeny



Didelot, Fraser, Gardy, Colijn MBE 2017

TransPhylo: https://github.com/xavierdidelot/TransPhylo

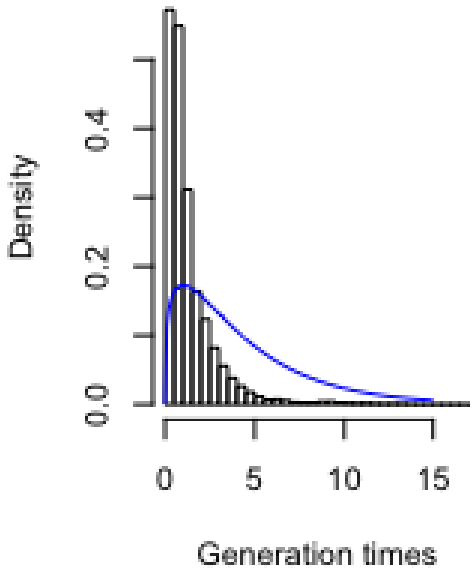
A 13-year TB outbreak in Hamburg

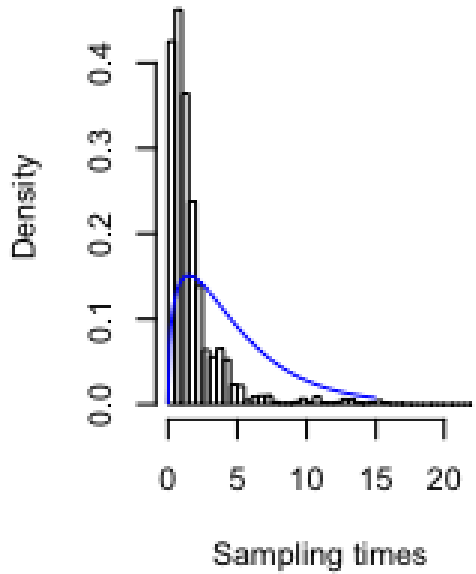
- Outbreak of 86 tuberculosis cases over 13 years. Roetzer et al 2013.
- Active case finding among contacts of cases
- Cases also identified for reasons other than TB infection
- Generation time and sampling time priors reflect uncertainty

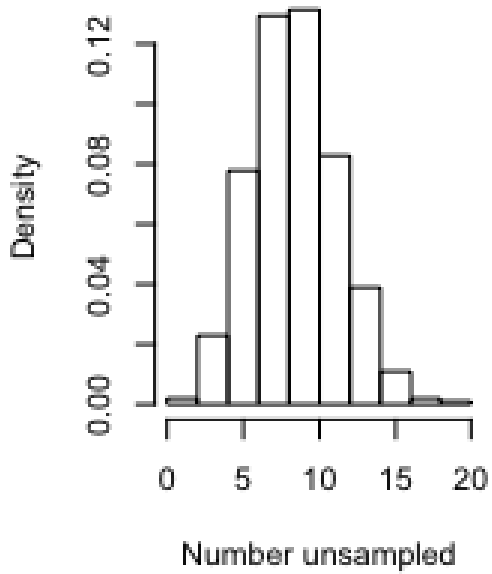
Time to infection, time to sampling, number unsampled

Cases infected someone within 2 years (80%) *among transmitting cases*. 75% sampled in 2 years.

It is likely that some cases were unsampled.







Prior in blue

When was each case infected?

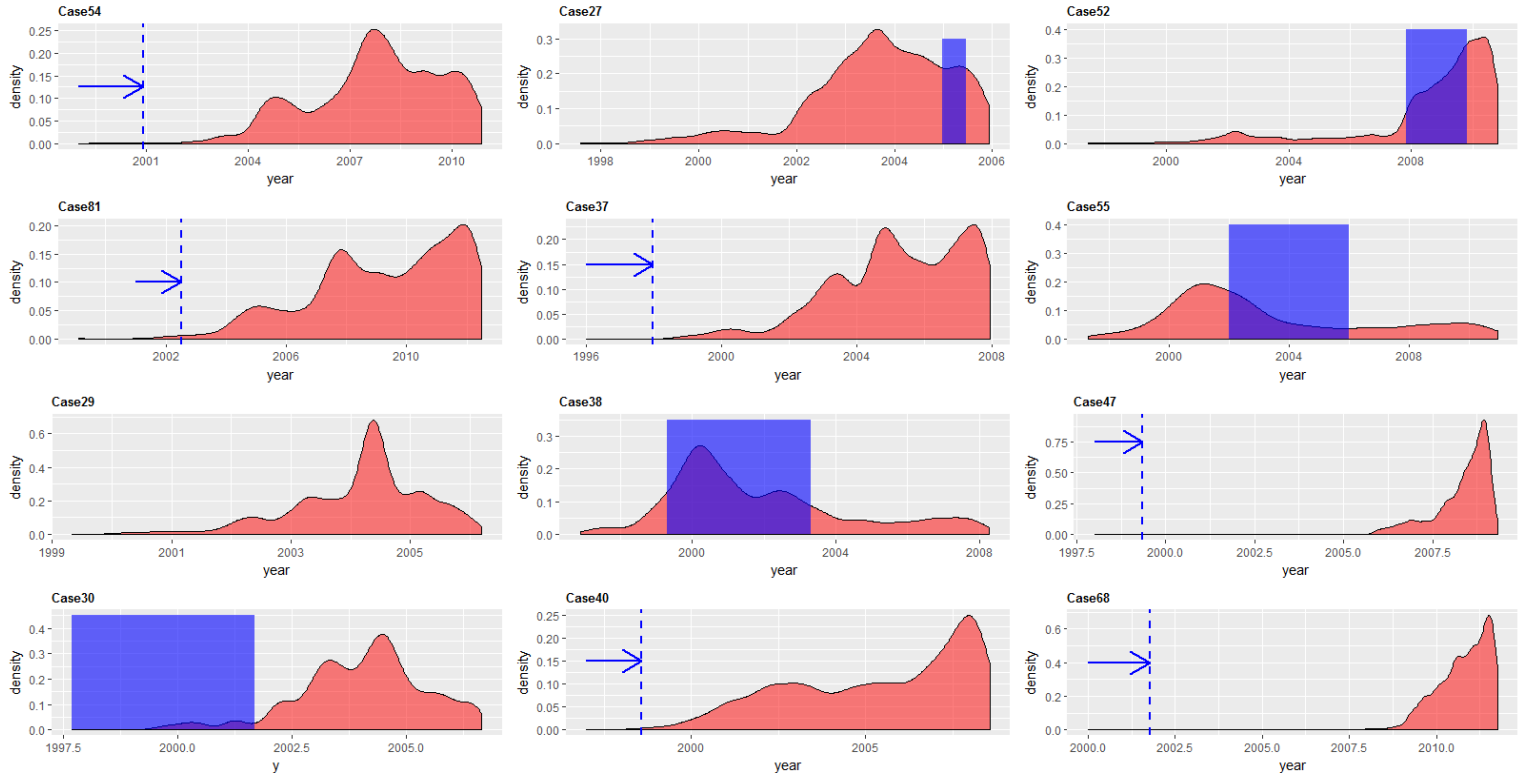
TransPhylo combines the priors for generation and sampling time, plus the genetic data, to give posterior times of infection for each case.

Data:

- Cluster of closely-related cases in Norway
- Cases occur among people immigrating to Norway
- It is often assumed that they were infected before arriving
- But genomic data show signs of recent transmission

We compared time of arrival to posterior time of infection for 13 closely-related cases

Posterior infection time compared to arrival in Norway



Red: Posterior time of infection. Blue: arrival in Norway.

Some cases were very likely infected in Norway.

Joint inference on many input trees

- Cluster data in low-incidence countries: many small clusters, large genetic distances away
- TransPhylo is likely to put many unsampled cases on the long branches and will not explore transmissions on clusters efficiently.
- Or - many input phylogenetic trees from a posterior collection
- This helps deal with tree uncertainty

Key improvement: use TransPhylo on *lots of trees* at the same time, *sharing parameters* between them.

Joint transmission inference – formal framework

- Start with lots of trees T_i : $T = (T_1, \ldots, T_n)$
- Likelihood
  - no parameter sharing - need to estimate all the θ_i:

    $p(T|\theta) = \prod_{i=1}^{n} p(T_i|\theta_1, \theta_2, \theta_3, ...\theta_n)$,
  - with parameter sharing - need to estimate only θ:

    $p(T|\theta) = \prod_{i=1}^{n} p(T_i|\theta)$

  where θ is one set of parameters common for all clusters.

- We can choose to share some or all of the parameters: analyse T_i simultaneously
- Estimates 3 parameters instead of $3n$. Each cluster is informed by the others.
- Computational efficiency, tighter CIs

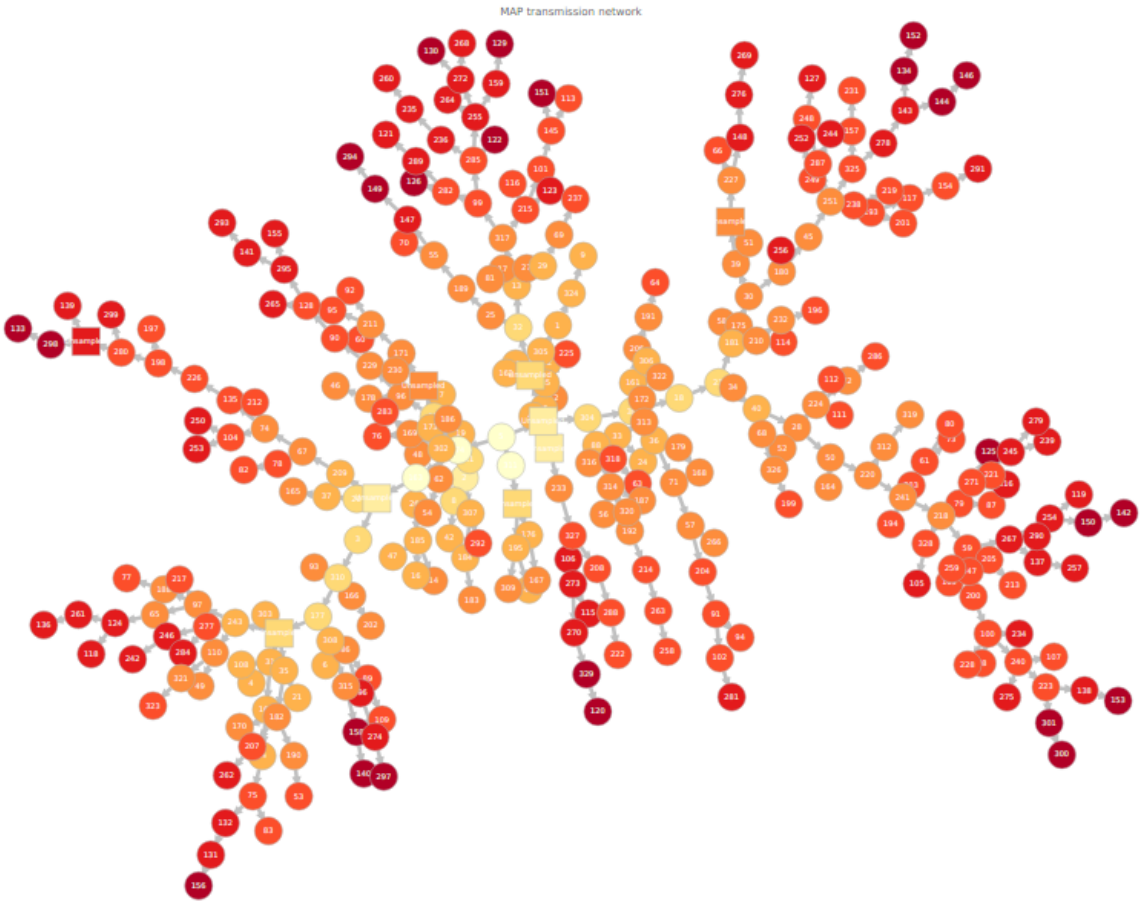
Long-term TB outbreak in London

- More than 400 cases over 20 years
- 351 sequences, 94 identical (!)
- Cases share a 24-locus MIRU type and isoniazid mono-resistance
- Outbreak showed signs of high transmissibility
- Previous analysis concluded that we can't get much from sequence data.

We re-analysed this outbreak with TransPhylo, using many BEAST posterior trees as inputs. This allows for uncertainty in the phylo tree.

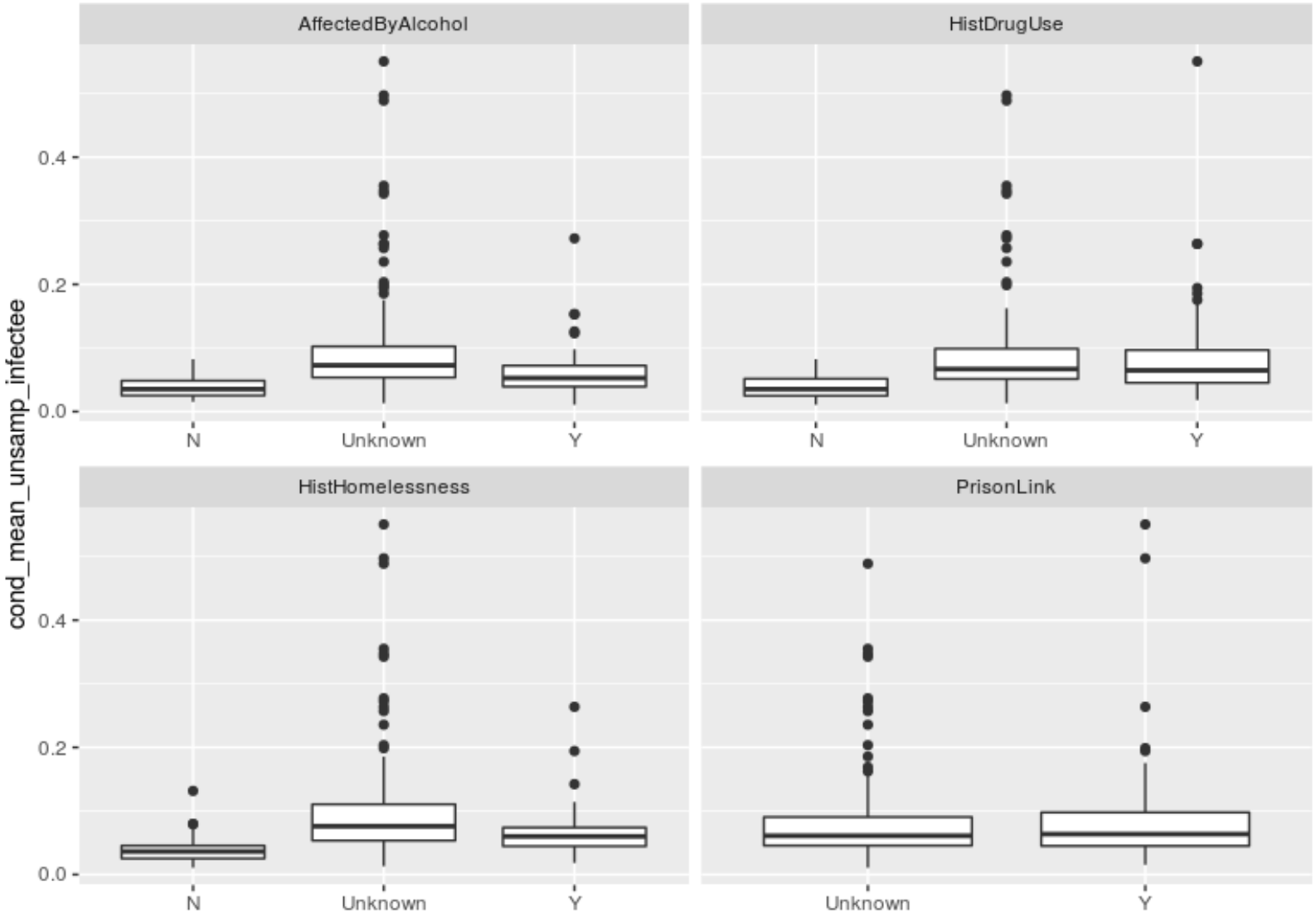
We related the TransPhylo outputs to clinical, demographic and individual data.

Reconstructed MAP transmission tree

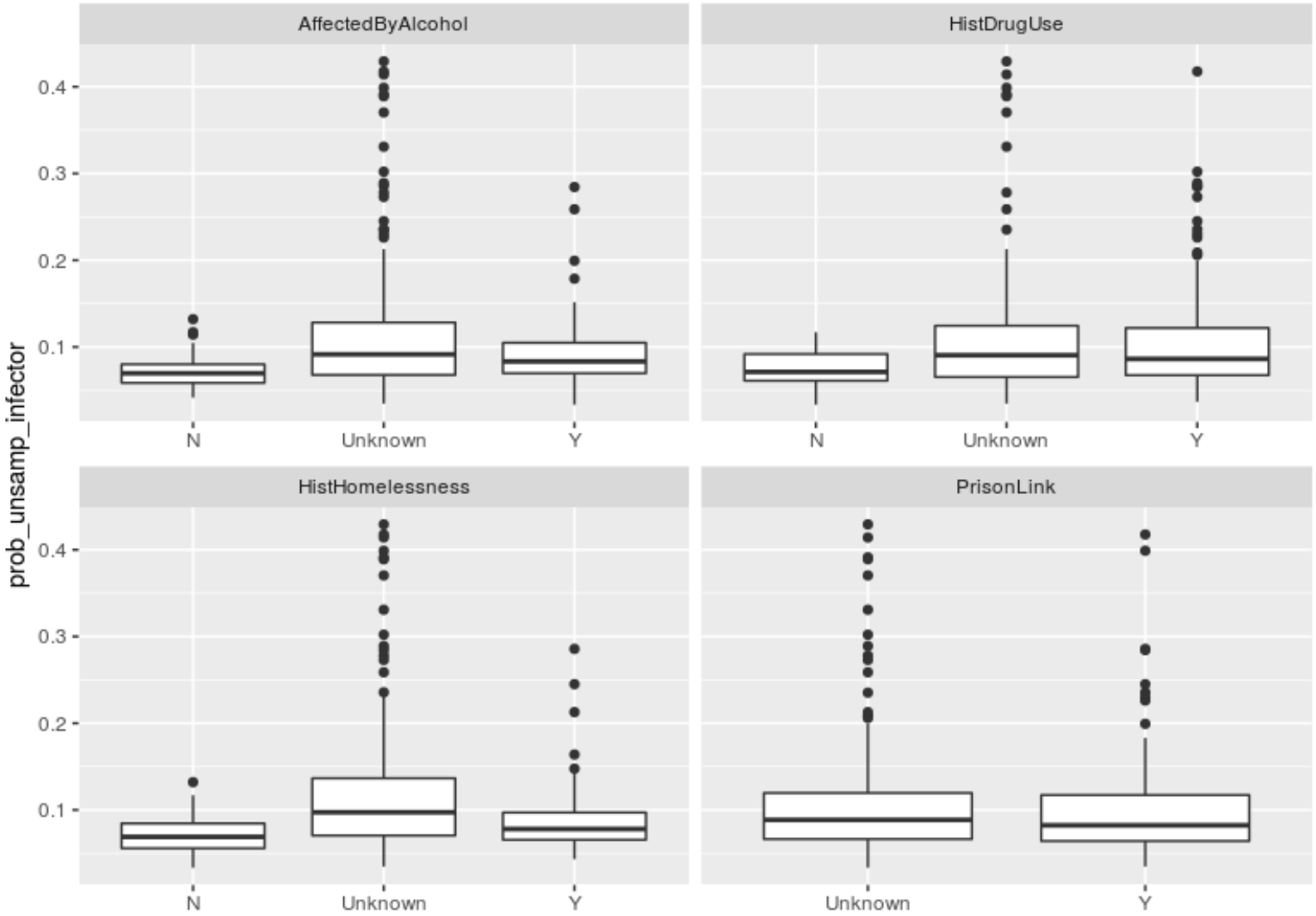


Colour: time of sampling. Shape: sampled (circles) or unsampled (squares)

Who tends to infect unsampled cases



Who tends to be infected by unsampled cases?

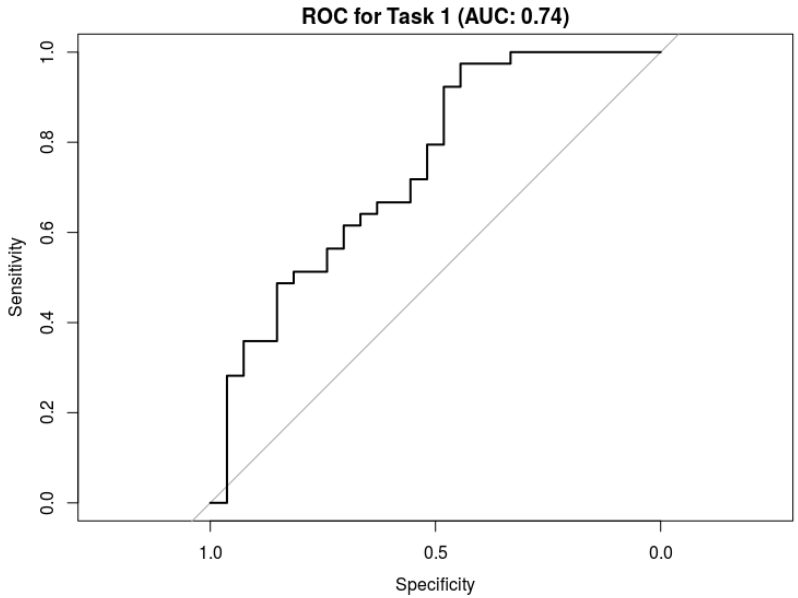


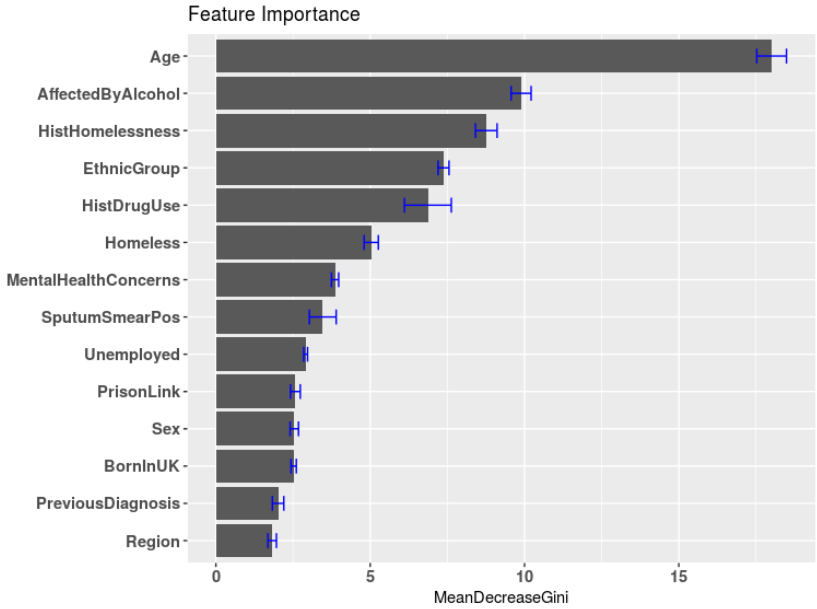
Machine learning to connect covariates to transmission

- Let "transmitter" status be: does the case infect others, in TransPhylo estimated trees, some fraction of the time?
- eg in at least 50% of the posterior trees
- This way, we estimate whether each individual host was a credible TB transmitter
- Note: there is no ground truth
- We have data on: alcohol, prison, drug history; age, ethnicity, London hospital location, smear status, and more

We train a random forest classifier to use our *other* data to predict whether a case is a credible TB transmitter.

Predicting credible TB transmitters from patient data





Age, history of being affected by alcohol, history of homelessness are the most important variables in predicting credible TB transmitters (in a random forest model)

R packages: TransPhylo, treespace

- `TransPhylo`: https://github.com/xavierdidelot/TransPhylo
  - New Rcpp implementation for the slowest part (Yuanwei Xu)
  - Module for outbreaker2
  - Xavier Didelot, Yuanwei Xu
- `treespace`: https://cran.r-project.org/web/packages/treespace/index.html
  - Metric comparisons for rooted, labelled phylogenetic trees
  - Michelle Kendall
  - Identify clusters (multi-modality) in your BEAST trees
  - Find median trees to represent those clusters - can be better than maximum clade credibility if trees are discordant

Thank you. Questions?

- James Stimson, Yuanwei Xu (TransPhylo extensions)
- Xavier Didelot, Christophe Fraser, Jennifer Gardy (TransPhylo)
- Vegard Eldhold (Norway data)
- Hollie-Ann Hatherell, Ibrahim Abubakar and Public Health England (PHE data)