Hindcasting Endemic Disease, with an example of FMD in Cameroon

Gustaf Rydevik

25 July 2016

Hindcasting FMD incidence in Cameroon using surveillance data with multiple diagnostics

Abstract

This paper discuss how cross-sectional data from multiple diagnostic tests with different temporal characteristics to hindcast times since incursion and epidemiological patterns of an infectious disease in an endemic setting. Assuming a cross-sectional sample of individuals infected with a pathogen and information on how the expected result of each diagnostic test varies following infection, a Bayesian MCMC approach is used for estimating time of exposure. We demonstrate this approach on a dataset of endemic FMD infection in randomly sampled herds in Cameroon. In addition, we use simulation approaches to demonstrate that our approach candistinguish between decreasing and non-decreasing endemic trends. Finally, we discuss the benefits of this novel methodology for the management of infectious diseases, and for evaluation of policy interventions.

Introduction

-State of the research of exploiting within-host processes -Copy thesis chapter here, update, and add new content

Pathogens are one of the major contributors to the burden of disease in humans[@Lopez2006], have a substantial economic impact on the livestock industry[@Stott2003], and can be a serious threat to conservation and management of wildlife populations[@Daszak2000]. A crucial component of efforts to control endemic disease is the use of infectious disease surveillance for tracking trends and evaluating the effect of control measures. The current state of human disease surveillance has been characterized as deficient in terms of both coverage and reporting speed[@Butler2006]. The more complex settings typical of livestock and particularly wildlife systems tend to result in the available surveillance data being sparser still for animal disease [@TMorner2002; @Perez2011; @TheRoyalSociety2002]

The structure of a functioning disease surveillance system is complex, with a string of tasks that need to be accomplished before a case is recorded in a database and becomes available to epidemiologists and policy makers. However, a crucial part is the use of diagnostic tests to identify and confirm the type of pathogen that caused infection. This and the following chapter will argue that combining two or more diagnostic tests with quantitative measurements of the force of infection provide substantial additional information that can be used to estimate historic patterns of infections. Current analysis typically do not make use of such information. Therefore, a novel statistical approach is introduced for recovering population-level trends of exposure even from only cross-sectional data by combining knowledge of the dynamic characteristics of multiple diagnostic tests to infer the timing of exposure events for individuals. The process of recovering such trends will be referred to as “hindcasting”, following terminology established in other papers [@Wethey2008; @Banakar2011; @Kleczkowski2007d] for reconstructing historical trends from currently available data. This chapter will focus on the potential use of hindcasting in the case of endemic diseases, while chapter four describes the potential for hindcasting in an epidemic setting.

Changes in the epidemiology and/or incidence of endemic pathogens are ideally tracked through the use of routine, ongoing surveillance. However, in a number of situations and for a number of pathogens, such ongoing surveillance is either non-existent or limited in its ability to provide a full, unbiased view. For some diseases, the epidemiology is known and the disease is considered important, but surveillance relies on diagnostic measures which are either expensive and underutilized, or lacking in sensitivity and/or specificity. One such example is the disease Scrapie in Sheep. Scrapie is a prion-spread disease with a very long incubation period, and difficult-to-detect symptoms. In the USA, Scrapie has decreased from a 0.2% prevalence to less than 0.05% between 2003 and 2009 thanks to introduced policy measures[@UnitedStatesDepartmentofagriculture2010]. However, there are substantial biases in reported prevalence numbers, raising the need for additional surveillance measures [@DelRioVilas2010]. Another pathogen, endemic in most of Europe, with similar characteristics is *Mycobacterium avium* subs. *Paratuberculosis*, also known as "Johne's disease". Paratuberculosis infections are asymptomatic for a long period of time, only detectable after some period and with the use of specifically targeted tests[ref!]. Reported prevalences across Europe vary widely, from 0.1% to 20%, largely owing to the difficulty of diagnosis [@Nielsen2009]. With these kinds of so-called iceberg disease systems, where routine surveillance only captures a small proportion of actual cases, there is a strong need for alternative strategies that can ensure that the trend measured by routine surveillance systems is representative of the full epidemiology of the targeted disease system.

There are a number of endemic diseases that are considered low importance and therefore not targeted by surveillance. When such a disease suddenly gains importance, because of increasing prevalence induced by changes such as mutation of the pathogen, or due to realizations of the extent of its economic impact, the ability to rapidly gain an understanding of the historic trends would be extremely useful in prioritizing and targeting interventions. This can be the case even for high profile pathogens such as the H5N1 flu virus, where the threat of silent spread in poultry flocks is a serious concern [@Savill2006]. Some pathogens have a very high incidence of undiagnosed infections, where the pathogen circulate widely in the population and causing non-specific disease. Salmonella[@Simonsen2011] and Pertussis[@Hallander2009a] are two examples of human diseases where the true extent of infections have been unknown until fairly recently. Sexually transmitted infections are also often under diagnosed because of social stigma associated with testing. Chlamydia is a disease with a significant disease burden in most parts of the world[@WHO2012b], and where the prevalence in women is much better known, and often reported to be higher, than in men for whom the testing rate is much lower(see e.g. the introduction of @Gotz).

For many endemic diseases, policies are put in place to reduce incidence or eradicate the disease - either locally, as with bovine viral diarrhea(BVD) in Scandinavia[@Stahl]; or globally, as happened with Rinderpest in Cattle[@FAO2013] and smallpox in humans[@WHO1980]. Measuring the impact of implementation of such policies is needed to ensure that eradication efforts are on the right track. High costs restrict the implementation of longitudinal surveillance programmes whereas cross-sectional studies of disease are more common. Therefore, methodology that could infer temporal trends from cross sectional data would be extremely beneficial. The application of the hindcasting techniques described here could be used to extend the utility of such cross-sectional studies to fulfill some of the objectives of an ongoing surveillance system.

Several papers have recovered limited historical characteristics of the spread of pathogens from cross-sectional data using a single diagnostic test, e.g. an antibody test. For example, Giorgi et al. estimated the time of the start of an HIV outbreak under assumptions of exponential growth of viral load [@Giorgi2010b]. Others have exploited information on diagnostic test kinetics, i.e., the pattern of diagnostic test values during the course of infection, to estimate average incidence rates. Example includes the use of antibody test kinetics to estimate sero-incidence rates for influenza [@Baguelin2011], salmonella in cattle [@Nielsen2011] and salmonella in humans [@Simonsen2008]. One challenge in these kind of studies is that the relationship between the magnitude of signals from diagnostic tests and time since exposure is usually not monotonic; the signals tends to increase and then decrease. This means that the inverse problem of estimating time since exposure given a test value is non-unique and although this can be framed as a statistical problem the resulting inference is highly uncertain [@Giorgi2010b],[@Simonsen2009], limiting what can be estimated from test data. However, there are often several diagnostic tests available that target different aspects of the multi-faceted dynamic interaction between host and pathogen [@Casadevall2001], and would thus exhibit different test kinetics. That is, the profile of test responses, as a function of time since exposure, will differ depending the

underlying diagnostic used. This mean that, in principle, we can generate a unique signal for a given time since exposure by combining results of several diagnostic tests that respond on different time scales. Here, this fact is exploited to develop a more robust statistical approach for analyzing cross-sectional field data from two or more diagnostic tests. Empirical infection models that characterize test kinetics are used to infer the time since exposure for each individual. While there is large uncertainty in the estimated exposure time for each individual, the combined estimates from multiple individuals describes the overall population-level distribution of infection times, which can be used to estimate the overall trend of incidence.

In an endemic setting, trends of infection are often gradual, and can be approximated by a constant change per time unit (month, year, decade). The chosen approach in this chapter was thus to posit that the incidence follows a linear trend with some slope, and that the disease is common enough that reinfections cannot be ignored. Section 3.2 develops the statistical framework for hindcasting in general, while section 3.3 details the mathematical consequences of assuming a constant linear trend with reinfections on inference of the trend from cross-sectional data. Section 3.4 details the choice and implementation of test kinetics. Section 3.5 details results from applying the framework to data simulated under a range of different scenarios. Finally, section 3.6 discuss the implications of the results and the hindcasting framework.

Framework for hindcasting endemic disease

Statistical framework

The statistical framework used for hindcasting in this thesis assumes test data y_{nk} from multiple disease diagnostics indexed by $k = 1, \ldots, K$ on individuals $i = 1, \ldots, N$. Each individual is assumed to have been tested at some time t_i, after having been exposed to the pathogen at some earlier time e_i. It is further assumed that these individuals are chosen in an unbiased, random manner from a larger population. Each diagnostic test is assumed to return a value in the form of a continuous 'level', which might, for example be the highest dilution at which antibodies are detected in a serological test. Without loss of generality, these levels are assumed to be scaled to the unit interval [0,1].

Initial exposure to a pathogen is the start of a complex dynamical process within the host. Such internal host-pathogen interactions can be conceptualized as a multivariate process that depends on the time since initial exposure. Each diagnostic test is assumed to target the state of a different component of this process so that each test k carried out at time t_i on individual i can be modelled as a latent variable $l_{ik}(t_i, e_i) = l_{ik}(d_i)$, with each test having differing but correlated response patterns over the time since initial exposure $d_i = t_i - e_i$. these latent variables are modelled using results from experimental infection studies for a given host-pathogen system, where the length of time since initial exposure d_i is known.

The known data, across all individuals in the sample, comprises a set of test results denoted by $Y = \{y_{ik}\}$ with sampling times $T = \{t_i\}$. The aim is to infer the unknown set of exposure times $E = \{e_i\}$, using information on the behaviour of the latent processes $L = L(T, E) = l_{ik}(t_i, e_i)$ generating the test results. In the hindcasting model, L represents the expected value of the test results given e_i and t_i. Note that when describing these sets the limits of each index $k = 1, \ldots, K$ and $n = 1, \ldots, N$ are implicit.

Under the hindcasting model, it is assumed that the sampling times T are precisely known whereas the quantities Y, L and E are assumed to be subject to uncertainty and variation. There are thus three components to the statistical model: a latent process model $P(L|T, E, \theta_L)$ describing uncertainty and variation in the host-pathogen interaction process within the host in terms of the time since initial exposure; a testing or observation model $P(Y|L, \theta_Y)$ describing the distribution of results from tests carried out on the hosts conditional on the internal latent process; and an epidemic trend model $P(E|T, \theta_e)$, describing the historical development of the epidemic in terms of the distribution of exposure times in the sampled host population, at the time of sampling. The specific implementations of each of these components in the linear trend setting is described in the next two sections. Combining the three parts of the model, the full data likelihood given an observed data set $\{Y, T\}$ is written as $P(Y, E, L|T, \theta) = P(Y|L, \theta_Y)P(L|T, E, \theta_L)P(E|T, \theta_E)$, where $\theta = \{\theta_Y, \theta_L, \theta_E\}$ Thus the likelihood combines models for testing with those for within and between host pathogen interactions.

According to Bayes' theorem, the so-called posterior distribution for the unknown parameters is proportional to the data likelihood and prior $P(\theta)$. Using the parameters of interest θ, the latent process L, the exposure times E, given the observed test data Y and sampling times T, the posterior distribution can be described by the equation

$$P(L, E, \theta|Y, T) = (P(Y, E, L|T, \theta)P(\theta))/(P(Y, T))$$

Within the Bayesian framework all inference is based on the posterior. The prior $P(\theta)$ can result from previous measurements or expert opinion, and represents knowledge about the values of the parameters before any of the data used in the likelihood is observed.

In what follows, the simplifying assumption will be made that the latent process L is modelled by a known deterministic function of T and E. This means that the term $P(L|T, E, \theta_L)$ drops out of the likelihood which then simplifies to $P(Y, E|T, \theta) = P(Y|L(T, E), \theta_Y)P(E|\theta_E)$, and the posterior becomes $P(E, \theta|Y, T) = P(Y, E|T, \theta)P(\theta)/(P(Y, T))$

Note that under this notation any parameters defining the deterministic latent process $L(T, E) = l_n k(t_n, e_n)$ are suppressed since they are not inferred i.e. $\theta = \{\theta_Y, \theta_E\}$.

In both cases above the normalization factor $P(Y, T)$ is typically unknown and computationally expensive to calculate. However, standard Markov Chain Monte Carlo (MCMC) methods circumvent this problem and are able to generate samples from the posterior even though the normalization is unknown (see @Robert2011 for an interesting overview of the historical development of this approach).

Estimating trends from times-since-infection data

Distribution of times since infection (tsi) under linear trend

The most basic scenario used for hindcasting a disease trend represents an endemic disease, with cases occurring at a constant rate. Formally, this scenario can be defined by assuming that the entire population is exposed to a force of infection λ. For a random observed individual, the time since last infection is then distributed according to an exponential distribution with rate parameter λ, $P(t < T) = \int_{x=0}^{T} \lambda e^{-\lambda x} dx = 1 - \lambda e^{-\lambda T}$.

This basic scenario was then modified to a scenario where there the force of infection has been changing over time according to a linear trend, $\lambda = \alpha + \beta t$. However, this linear trend describes the incidence over time if the cases are reported continuously. If, instead, the cases are collected at a single point in time at some point after infection, the distribution of interest is then the distribution of times since last infection("tsi") in the population, hereafter denoted $f_{tsi}(t)$, under the assumption of a linear trend.

The incidence of infection can also be referred to as the average *hazard rate* of infection. Given a constant hazard rate *lambda*, the probability of the event $I% of having been infected by time t is given by $1 - e^{-\lambda t}$. In the linear trend scenario, $\lambda = \alpha + \beta t$ as mentioned above. Because of the linearity of the trend, over a time period from 0 to t, the probability of having been infected before that time is equivalent to the probability of having been infected under a constant trend of the mean incidence over the period, $\hat{\lambda} = \alpha + \beta t/2$, and so by analogue to the constant case, this probability can be written as $P(I < t) = 1 - e^{-(\alpha+\beta*t/2)*t}$. From this, the probability density function for the times since infection can be calculated as $f_{tsi}(t) = d(P(i < t))/dt = (\alpha + \beta t)e^{-(\alpha+\beta t/2)t}$.

In the implementation, this distribution was assumed to be censored at some time point in the past C, and it was further assumed that it was possible a priori to distinguish individuals that had been infected at some point during this time period, from naive individuals. When implementing such a censoring, the equation above needs to be modified by an additional scaling factor $1 - e^{-(\alpha+\beta C/2)C}$, equal to the integral of $p(t)$ over the time span $(0, C)$. The full equation used to represent the distribution of exposure times was thus

$$p(t) = (\alpha + \beta t)e^{-(\alpha+\beta t/2)t}/(1 - e^{-(\alpha+\beta C/2)C})$$

As the model is implemented in the Bayesian framework, priors for both the incidence (α and trend (β) parameters needs to be specified. In order to provide a prior for the incidence, information about the

population size needs to be incorporated. This was done by noting that the number of positive and negative individuals in a population can be approximately described by a binomial model, parameterized by the probability of infection p. Then denote by N_+ the number of positive individuals known to have been infected during a time period C, from a population of size N. With an uninformative $Beta(1,1)$ prior for the probability of infection p the distribution of the probability of infection given the number of positives and negatives observed is $p \sim \beta(N_+ + 1, N - N_+ + 1)$. Under the assumption of a linear trend, the mean incidence over the time period C is equal to the the incidence at time $C/2$, $\hat{\lambda} = \alpha + \beta \times C/2$. The proportion p of observed positive individuals are exactly one minus those that had not been infected during any of the time periods up until the time of censoring C. From this observation, the mean incidence $\bar{\lambda}$ per time unit can be derived from the proportion of positive individuals over the time period C by the relationship:

$$p = 1 - (1 - \bar{\lambda})^C \rightarrow 1 - \bar{\lambda} = (1-p)^{1/C} \rightarrow \bar{\lambda} = 1 - (1-p)^{1/C}$$

For the trend parameter β, note that if the linear trend model is assumed to hold over the time period C, then the incidence is not allowed to become negative over this time. Using this, the restriction for the trend becomes:

$$\bar{\lambda} \pm \beta \times C/2 > 0 \rightarrow$$
$$\bar{\lambda} > \beta \times C/2 > -\bar{\lambda} \rightarrow$$
$$\bar{\lambda} 2/C > \beta > -\bar{\lambda} \times 2/C$$

Following this, the trend was assigned a uniform prior, $\beta \sim U(-\bar{\lambda} \times 2/C, \bar{\lambda} \times 2/C)$. The intercept parameter α was then simply calculated from trend and $\bar{\lambda}$ via $\alpha = \bar{\lambda} - \beta \times C/2$.

At this point, it should be pointed out that the properties of distribution $f_{tsi}(t)$ of times since infection are somewhat counterintuitive. Figure 3.1 shows its shape for decreasing ($\alpha = 0.05$), constant ($\alpha = 0$), and increasing ($\alpha = -0.05$) parameter values, holding β constant to 0.1. The first thing to note is that because we are looking backwards in time, coefficients have opposite sign - $\alpha = 0.05$ denotes that the incidence rate has been decreasing by 0.05 per time unit, whereas $\alpha = -0.05$ denotes that the incidence rate is increasing. The second thing to note is that all three curves have the same upwards slope. The further back in time we look, the less likely we are to find a case that occurred at that time. In effect, by only considering the time since last infection, we are assuming that a reinfection resets the "clock" of the infectious disease dynamics. This in turn means that more recent infections can hide infections that occurred further back in time. However, by looking at the curvature of the exponential distribution, it is still positive to estimate the actual incidence trend.

Properties of the linear-trend induced distibution of times since infection

The aim of this chapter is to recover the population-level trend of incidence from test measurements taken from individuals that have been infected at a some point in the past in a population where the time-since-infection (tsi) distribution is defined above, i.e.

$$f_{tsi}(t) = (\alpha + \beta t) e^{-(\alpha + \beta t/2)t} / (1 - e^{-(\alpha + \beta C/2)C})$$

[number equation!] In order to study the properties of this inference problem, a first approach is to investigate the simplified situation where the times of infection are known and generated using f_{tsi}.

Including the priors described in the previous section, the expression for the log likelihood of f_{tsi} given observations of time since infections X (denoted by LL_f), becomes

$$LL_f(\alpha, \beta | X) = \log(U(\alpha | -2 \times \bar{\lambda}/C, 2 \times \bar{\lambda}/C)) +$$
$$\log \beta(\bar{\lambda} | N + 1, N - N_+ + 1)) +$$

$$\sum_{\forall i} log[(\lambda + \beta \times X_i) \frac{e^{-(\lambda+\beta\times X_i/2)X_i}}{(1 - e^{-(\lambda+\beta\times C./2)\times C))}} \times I(X < C)]$$

Times of infections $X = \{x_i\}$ were simulated from the probability distribution of $f + tsi$, and the value of the log likelihood LL_f given the simulated data was calculated over a grid of values for α and β. Typical results are shown in Figure 3.2. Note that this is assuming that the times of infection were exactly known.

Figure 3.2 shows the resulting likelihood surface of LL_f. There are two things to note in this image: the first one is that the region of highest likelihood is in a region surrounding the black line. This black line is the line for which the combination of α and β results in the same average incidence $\bar{\lambda}$, which indicates that the Beta prior on $\bar{\lambda}$ has a strong influence on the curvature of the likelihood surface. The second thing to note is that along the line of equal mean incidence, there is little change in the colour, and that there is little distinguishing the region surrounding the true value (noted by the black dot), from the rest of the region. This indicates that while it will be relatively easy to recover the value of $\bar{\lambda}$, finding the correct combination of α and β is more challenging.

Sampling from the posterior using Stan

For conducting inference of the endemic trend, the posterior distribution of parameters described in the previous section is evaluated, conditional on observed test data and knowledge of expected test kinetics.

A high level language for hierarchical Bayesian models known as JAGS [@Plummer2003] was used to implement the statistical framework (see appendix for example JAGS code) and evaluate the posterior distribution using the Metropolis-Hastings algorithm combined with Gibbs sampling (see section 1.X for a more detailed discussion). The code was called from within R using the *rjags* package [@Plummer2014]. Samples were taken from the posterior distribution of time since infection for each individual, as well as the posterior distributions of the parameters of the trend of incidence.

As noted in the introduction (section 1.X) a key question with the implementation of MCMC algorithms is that of convergence and mixing. The reliability of our sampling tools were assessed using trace plots. Figure 3.5 shows example trace plots from three scenarios (decreasing, constant and increasing trend) of the last 1000 draws (thinned so that every 10 draw is shown) of the population level trend parameter β, the intercept parameter α, and the mean incidence $\hat{\lambda}$ (see equation XX), from five different chains after all chains have been run for a 1500 iteration burn-in. The chains mixed well for all scenarios, with Gelman-Rubin statistics of ~ 1.00 for all three scenarios and parameters (for a more detailed discussion on issues of convergence, see section 1.X).

For the full range of scenarios, it was not feasible to inspect trace plots. Instead, a Gelman-Rubin statistic above 1.5 was used to filter out those runs that had not converged (~5% of the total runs).

Example: Hindcasting FMD infections in Cameroon

Background

We here demonstrate an approach to estimating the time since last incursion of Foot-and-Mouth disease virus(FMDV) for individual herds, in an endemic, low income setting. The data set we analyze covers 1500 cattle distributed across XX randomly sampled cattle herds from the YY area in Cameroon. This data has been extensively described elsewhere [1,2,3,...]. The cattle were tested for the presence of FMDV using a range of diagnostic approaches, including antibody tests, virus culture, and clinical signs. In addition to diagnostics tests, the herdsmen was asked to provide estimates for the time since last incursion of FMDV. The herdsmen estimates has been shown to be broadly reliable[ref]. In the following sections, we will demonstrate how we developed a model for predicting the herdsman estimates from diagnostic test results. This model can thus be used to hindcast the time since incursion for a given herd, and estimates can be combined to provide overall estimates of spatial patterns or trends of FMD in the YY district of Cameroon.

Data

The data used was collected by Bronsvoort etal in Cameroon. Full details can be found in ....

There was the following diagnostic measures used on included cattle:

-Danish C-ELISA FOr non-structural protein, so will pick up any serotype in theory.

kinetics from Bronsvoort2004

-South American I-ELISA Also non-structural and will pick up any serotype

-CHEKIT kit Final non-structural (least good, "sensitivity" of ~23%)

kinetics from Bronsvoort2004

-EITB - enzyme linked immunoelectrotransfer blot Non structural, but a binary yes/no response. Americans use EITB and I-ELISA as a combined diagnostic

FMD_VNT are virus neutralisation test, where you do sequential dilutions of the virus and look whether antibodies react or not.

Higher the number, the greater the dilution with detectable virus levels. FMD has seven different serotypes, with different geographical distributions. Three VNT tests: A-serotype, O-serotype and SAT2 ("south-african territories") serotype.

PRobang is used to scrape the cells to collect viruses for cultivation. Only done on animals where we think that something is going on. That's PbP A,o,and Sat2. From herds 32 onwards, we have collected the probang. The numbers are (probably) binary classifiers based on Antigen ELISA-results.

Finally, FMDS O, A, and SAT2 are the binary classifiers based on the VNT tests.

FMD_VNT_A

FMD_VNT_O

FMD_VNT_SAT2

If an animal gets exposed, the animal will develop a particular antibody response that will last for years. The VNT results will therefore likely remain for a long time following exposure/infection. The non-structural ELISA results will disappear within 6-12 months. The probang has some sort of exponential decay of viruses/ likelihood of cultivation. So a latent probability of some sort...

An additional thing for the future is that we have recording based on the growth of hoofs and how far up there are records of old lesions.

In terms of animals, we have clinically infected, recently infected, old lesions, and healthy animals.

Monlast is the herdsmans reporting, and can be used as a validation to compare with estimated times since infection.

One of the practical questions is to go through a herd and try and age lesions. But they did run into problems with sheep flocks that they missed in 2001, since sheep tends to not have clnical signs. Using a stat method would be useful here - something to bring up in a discussion. In pig farms, at least three infection cycles

before the outbreak was detected. Want to try and identify how far back was the herd infected. Being able3 to do this quite quickly would be very useful.

We can use Alexanderssen2003 as how a VNT response would look like post infection, but obviously conditional on a scaling factor to account for different units. Would go off the contact one.

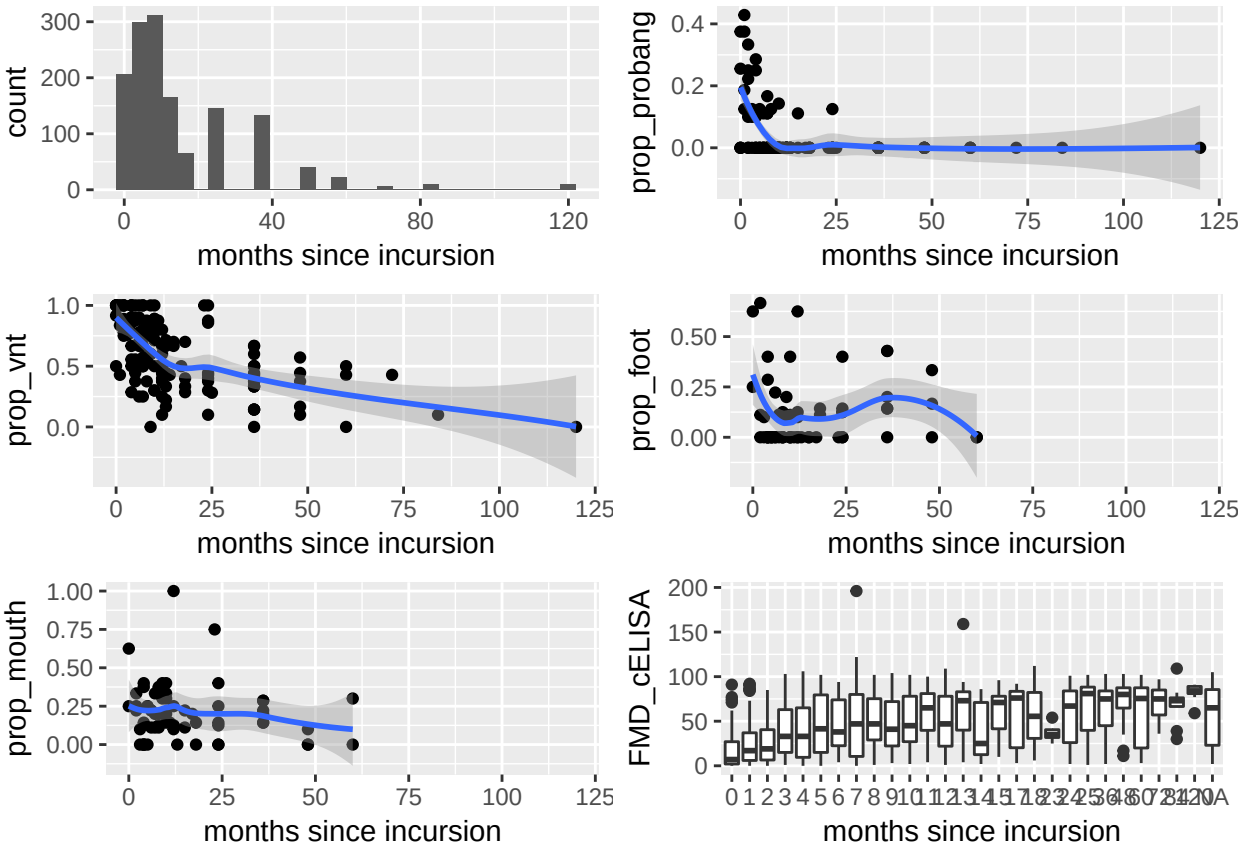
Names of researchers:

Bergmann (iELISA) Bronsvoort Sorensen KJ (cELISA) Alexandersen Hamblin Brocchi Cuncliffe Dekker (review!)

People doing long-term experimental infection/post-disease follow up. Possibly Bergmann Pirbright does short, 1-month studies.

Section 1: using the cameroon data to estimate time since incursion.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 27 rows containing non-finite values (stat_bin).
## `geom_smooth()` using method = 'loess'
## Warning: Removed 29 rows containing non-finite values (stat_smooth).
## Warning: Removed 29 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess'
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## Warning: Removed 3 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess'
## Warning: Removed 87 rows containing non-finite values (stat_smooth).
## Warning: Removed 87 rows containing missing values (geom_point).
## `geom_smooth()` using method = 'loess'
## Warning: Removed 87 rows containing non-finite values (stat_smooth).
## Warning: Removed 87 rows containing missing values (geom_point).
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



We have diagnostic data at the individual cow level; Probang measurements, VNT measurements, and ELISA measurements: $y_{i1} = Probang_i$ (which is binary) $y_{i2} = VNT_i$ (which could be treated as binary or as continous) $y_{i3} = Lesion_i$ (which is ordinal).

These are all measurements of three different latent processes. The probang results measures whether or not it was positive to successfully culture FMDV from the throat(?) of the animal, and can thus be seen as an indirect indicator of the amount of virus present in the throat. This is the first latent component, $l_{i1}(d_i) = exp(d_i, \theta_{probang})$. We here assume that the the virus declines exponentially following infection, which is a simplistic but useful approximation.

The VNT results is an indicator of the level of FMD-specific antibodies present in the blood of the animal. This is the second latent component, $l_{i2}(d_i) = Logistic(d_i, \theta_{VNT})$. In this case we assume that the antibody response follows a logistic growth curve.

Finally, the lesion aging attempts to measure the amount of new growth that has occured on the hoves of animals since an FMD infection (which leaves very characteristic markings at the join between legs and hooves,that then migrate downwards). $l_{i3}(d_i) = d_i * \theta_{lesions}$. We will simply model this process as a linear function of time since infection.

The full latent process can thus be written as

$P(L|T, E, \theta_L) = exp(d_i, \theta_{probang}), Logistic(d_i, \theta_{VNT}), d_i * \theta_{lesions})$

$monlast\ prior(f(incidence))$

$P(probang - positive, VNT - Positive)\ f(monlast, age, ...)$ ## Could separate, or not... $VNT \sim f(monlast)\ ELISA \sim F(monlast)$

$P(probang = 1) = Bernoulli(logit(\theta_1))\ \theta = f_1(T_{infected}) + g_1(age, age^2))$

$P(Vnt = 1) = Bernoulli(logit(\theta_2))\ \theta = f_2(T_{infected}) + g_2(age, age^2))$

Analyses of the cameroon test data

Simulations of different timescales and types of diseases

Should we include this? if so, copy from thesis ## Discussion (copy from Thesis, then update, and add new content.)

This chapter has introduced and tested a novel technique for hindcasting the history of exposure to disease in a population using only cross-sectional data combined with information on pathogen test kinetics. The results demonstrate that this procedure enable the estimation changes in disease incidence over time. The results also demonstrated how this approach is able to distinguish between an increasing trend and a stable, or decreasing trend, as well as produce posterior estimates quantifying this disease trend. This goes beyond previous sero-incidence studies which estimated the average incidence in a population, without attempting to estimate temporal trends in prevalence. [refs here!]

The use of Lotka-Volterra (LV) equations to describe the pathogen-host dynamic, and thus the paired-test development over time, made it possible to consider several different archetypes for the pathogen-host dynamic. Hindcasting was found to be possible for all of the different archetypes examined. Further, different archetypes proved to result in very similar overall hindcasting performance, with the exception of robustness to differing number of tests used. The result on number of tests versus performance in 3.6.3 seem to indicate that a combination of diagnostic tests are more robust across the board than a single diagnostic test; and that it may be possible to use a single diagnostic test for hindcasting, but it may also fail completely. This seem to warrant more research into the specific requirements of disease kinetics for hindcasting.

The results from evaluating scenarios with parameters close to observed epidemiological patterns in Scrapie, Chlamydia, and Squirrelpox, indicate that useful precision levels can be reached with realistic sample sizes.

In real world applications, the kinetics used to inform the hindcasting technique would likely be derived from other published data, such as experimental infection studies. In such cases, the LV calculations could be replaced with a simple lookup table for the expected mean response of the test at a given point in time, combined with information on the variability of the test. Alternatively models, such as the LV equations, fitted to the available data could be used.

The natural pairing of tests to model with the LV approach is a nucleic acid test for genetic material from the pathogen (e.g. a realtime PCR test), combined with a test measuring the antibody test response, such as a quantitative ELISA test. However, any type of paired tests commonly used for pathogen diagnostics could be used. Other examples are a pairing of a culture-based test combined with IGG antibodies, or even the severity of symptoms measured on an ordinal scale combined with viral load measurements. Thus, a wide range of diagnostic measures could potentially be used within the hindcasting framework presented here.

The results from this and the next chapter provides strong arguments in favor of recording the raw test results together with the resulting diagnosis, and for utilizing more than one diagnostic test whenever feasible. Thus, when setting up surveillance systems, it should be emphasised that the results of all diagnostic tests used should be recorded in the database. Such a database should also detail the quantitative level of evidence, in addition to the regular binary ”infected/non-infected” result. The cost of conducting and recording the result of two diagnostic tests should be considered in relation to the benefits. For example in terms of feedback to farmers and policymakers on the impact of control measures and for detecting any potential costly changes in the prevalence. It should also be noted that the methods introduced here enable such benefits to be derived from cross-sectional data and therefore the additional costs described above should also be compared with the costs of running longitudinal studies.

An obvious extension to the work presented here is to consider more complex changes in pathogen incidence than simple linear trends. In principle, since the hindcasting procedure provides approximate times of exposure any model that describe the pattern of times of exposure could be considered. The linear trends described here are primarily suitable for endemic diseases. Therefore, in the following chapter, the development of the hindcasting technique is continued by considering an outbreak of the pathogen in which the rise and fall

of the epidemic is adequately described, using a lognormal distribution of exposure times instead of linear trends to capture the rise-and-fall of epidemics.

References