Model to assess the usefulness of sequencing technology in detecting emerging antimicrobial surveillance and forecasting antimicrobial resistance in the animal primary production.

Deliverable D-JRP6-5.2 An assessment of the public health effects of very different surveillance strategies to detect emerging foodborne infections in a MS or at European level.

Ofoosuhene Okofrobour Apenteng DTU
Håkan Vigre DTU

Table of Contents

Summary

The overall aim of this work was to develop a modelling framework for a holistic assessment of how a new laboratory technology could be used to improve the performance of surveillance systems before the method is actually implemented. This case study simulated the use of metagenomics in the monitoring programme of antimicrobial resistance (AMR) in the Danish pig population. The assessment of the performance was both on detecting changes in the occurrence of existing AMR and to detect the emergence of previously undetected genes coding for specific AMR.

The ability to take action on the increased occurrence of AMR is dependent on a monitoring system of AMR in humans, animals and in the environment. In Denmark, the current surveillance programme DANMAP is based on isolation of indicator bacteria from samples in humans, animals and food. These samples are tested phenotypically for AMR using minimum inhibitory concentration (MIC) panels. Currently, only *E*. coli and *Enterobacteriaceae* are used for monitoring phenotypic occurrence of AMR in the microbiome in animal populations. This limits the detection of changes in AMR to only those genes that are expressed in these indicator bacteria. When AMR is emerging in a population, it might occur in one, few or several bacteria species, and this may not be immediately detectable with the current strategy.

Due to the rapid development of gene sequencing technology, it is now feasible to investigate the presence of all known AMR-genes in all bacteria in samples directly (metagenomics). Therefore, by the use of metagenomics, it may be possible to detect emerging AMR in a population earlier than with phenotypic test of indicator bacteria.

The model framework developed for this case study has two dimensions: 1) the dynamic change in the occurrence of AMR genes in the population over time and 2) the monitoring procedure (sampling schedule), laboratory method, interpretation and statistical analysis of laboratory results. The simulation of the monitoring procedure included modules describing stochastic processes along the whole chain from sampling including: random selection of herds and animals, pooling of samples and thereby dilution of the genes of interest, randomness in the sequencing procedure related to how much genetic material that is sequenced. Using this detailed modelling approach, it was possible to assess how much of the uncertainty in the surveillance programme that originated from the sampling procedure and how much originated from the diagnostic procedure. This information can support surveillance design that reduced uncertainty of the surveillance results.

The data obtained from the metagenomic procedure has a quantitative format in the form of the number of genes in a sample. It is therefore possible to utilise more computationally intensive methods in the analysis of the laboratory results. In this case study, we explored different statistical analyses to objectively detect changes in the observed data, and to forecast AMR based on the trend in the historical data. Forecasting is important for the control of AMR because precautionary actions need to be implemented early before the problem becomes a burden on human health.

In the design of many surveillance programmes, decision makers must make a subjective eye-ball forecast based on graphs illustrating the historical occurrence over time. The ability for risk managers to make evidence based decisions about how to control risks prospectively, depends on the information available. In this context objective forecasts based on the observed data are highly informative for supporting decisions. The importance of forecasting is of specific relevance for health issues as AMR,

because the time between implementation of control actions and effect on AMR is relatively long, on the scale of months or years.

1. Introduction

The ability to take action on the increased occurrence of antimicrobial resistance (AMR) is dependent on a monitoring system of AMR in humans, animals and in the environment. In Denmark, the current surveillance DANMAP is based on isolation of indicator bacteria from samples in humans, animals and food and phenotypic and tested for AMR using MIC panels. AMR can be expected to occur in all bacteria species. Using only a couple of indicator bacteria (E. coli and Enterobacteriaceae) for detecting emerging increased occurrence of AMR is therefore depending on that the AMR is present in the indicator bacteria. An increased occurrence of phenotypical resistant indicator bacteria can be due to both the environment favouring clones of the indicator bacteria that is carrying genes encoding for AMR, or that the indicator bacteria has picked up mobile genetic elements from other bacteria species that has been favoured by carrying genes for AMR. The sensitivity of the current system to detect emerging AMR is totally dependent on that the AMR is expressed in the indicator bacteria, which is absolutely not always the case in realty.

With the ongoing development of gene sequencing technology, it is now possible to investigate the presence of AMR in all bacteria in samples directly (metagenomics), using gene sequencing and subsequently mapping sequencing against databases with known sequences of nucleobases for AMR genes. The use of metagenomics increases the likelihood that emerging AMR is detected earlier due to that the whole microbiota of the sample is analysed for AMR, instead only selected indicators, and that the AMR does not need to be expressed by the bacteria in the laboratory.

However, the likelihood to detect an emerging gene in a population using metagenomics depends on a number of factors such as sample size and frequency, how many DNA fragments sequenced in the sample and the actual occurrence of the AMR gene in the population. The objective of the presented study was to perform simulation studies estimating the time to detecting emerge of AMR in a population of production animals in the country based on different sample plans and metagenomics. The study was performed by first simulating emerge of AMR in a population over time – we utilised real metagenomics data from the Danish pig production, and emerge was simulated using epidemiological models such as SIR. Secondly, a given number of faecal samples was collected stochastically according to a pre-defined sampling schedule. The sequencing depth of the metagenomics were included as a stochastic process. We run three different scenarios for emerging AMR. In all scenarios the time until detecting of emerge after the emerge was actually started was estimated. This was done using the DBEST (Jamali et al., 2015b) .

The aim with this work was to develop a framework for a holistic assessment of how new laboratory technologies can be used to improve the performance of surveillance systems before the technology is actually implementation in a population. The framework presented is developed focusing on how gene sequencing could be utilized in the monitoring program of AMR in the Danish pig population. In this monitoring program, focus is both on detecting changes in the occurrence of existing AMR, and to detect emerge of previously not present genes coding for specific genes.

The framework is assessing the performance focusing on one outcome per application of the framework. However, the use of gene sequencing of the whole matrix (metagenomics), instead of focusing on pre-defined bacteria species, makes the whole monitoring much broader, looking for all known AMR genes in each sample.

2. Materials and methods used in the case study

In this study, the approach was to first simulate data representing the "true occurrence" of AMR genes in a population of 4.000 farms producing slaughtering pigs. Secondly, we simulate the surveillance program running in the population, taking into account the stochasticity in sampling and laboratory procedure. The simulated "observed" results was the analysed statistically to detect changes and forecast the occurrence of AMR in the future. Finally, the result from the surveillance program was compared to the true occurrence, and the time to detect after emerge was estimated, and the accuracy of forecast was assessed.

2.1 Data

In the study, data about the actual occurrence of AMR genes in the animal population was simulated utilising measured amount of AMR genes in approx. 100 Danish pig farms producing slaughtering pigs. The occurrence of AMR genes were measured using metagenomics – the diagnostic procedure of interest in the case study.

2.2 The simulation models

In the simulation of the observed surveillance results, the results depend on:

- if it is individual samples or pooled samples that are analysed
- number of animals in the pool (dilution effect)
- the farm prevalence
- the within farm prevalence
- the concentration of the AMR in the animal
- the amount of the genetic material in the sample that is sequenced

These values are all stochastic in reality and in case of emerging AMR, the occurrence at farms and animals vary by time, and these sources of variation and changes was built into the simulation model using Monte Carlo simulation. The susceptible-infected (SI) model was used to simulate the spread between farms (Apenteng et al., 2020). The stochasticity in the gene sequencing was modelled using binomial distribution describing the random selection of which genetic material in the sample that is actually sequenced. All analyses were carried out using R software. Finally, the observed data was compared with the true data to estimate "time to detection after the true start of the emerging AMR in the population. Also, the observed data were analysed using a time series approach to find the trend within the observed data in order to i) detect the change in the occurrence of AMR genes within the population and to ii) forecast the occurrence AMR in the future.

2.3 Modelling true occurrence and emerge of AMR-genes

The true occurrence of different AMR genes in the population was defined with:
The farm prevalence; the within farm prevalence and; the concentration of the gene in the faeces of in infected pig. The concentration of the AMR gene(s) of interest in faeces was obtained from a Weibull distribution fitted to real-world data of the concentration of genes in pig faeces.

In the case study we work with three different true scenarios:

(1) An increased occurrence of a type of AMR resistance that is present in all farms in all pigs (endemic). In this scenario, we work with data representing the occurrence of tetracycline resistance in the pig population. The settings for the truth was that the prevalence of farms and animals is 100%. The initial concentration is equal to the concentration of tetracycline resistance genes observed in Danish pigs, and the increase is 5% per year.

(2) Same as (1), but for beta-lactam (resistance against penicillin, cephalosporin, etc). The concentration of these genes are much lower in pigs compared to tetracycline.

(3) The introduction of a new resistance gene into one farm, which will spread to other farms over time (emerging). The spread of the AMR gene between farms was modelled using a SIR approach, with a rate equal to a spread resulting in 80% contaminated farms after 5 years. In scenario 3, we assume that the concentration of the emerging gene was the same as the concentration of the blaOXA gene (a resistance gene with an extreme low concentration in the pig population).

All scenarios were simulated in a population of 4.000 farms.

2.4 Simulating the sampling schedule and concentration of genes in the sample used for gene sequencing

In the case study, we assume that the sampling was performed once per month, were faecal material were collected at the slaughtering line from a give number of pigs, and that all material was pooled into one sample that was sent to the laboratory.

The sampling procedure was performed randomly, at we run three different models varying number of samples that were pooled (*n* =, 5, 20 and 100). The concentration of a specific AMR gene in the pooled sample was a stochastic function of how many of the samples that were from infected pigs, and the actually concentration of the gene in the faeces of each pig carrying the gene. The effect of poling was modelled as estimating the mean concentration among the pooled samples.

In reality, observed data in a surveillance program just represent one out of an infinitive combination of observations obtained when analysing samples from the reality, we did 101 iterations for each scenario – representing 101 possible outcomes of a surveillance system.

In scenario 1 and 2, representing an increase of already present AMR. Each iteration represents a surveillance system running for 240 months, whereof the first 120 month represent a steady state in the population, and were the increase start at month 121.
In scenario 3, the emerging gene was seeded into one herd at month 1. Subsequently the gene spread across the population according to the SIR model. Each iteration was ran for 60 months, and at month 60, the gene of interest was present in almost the entire population.

The stochastic part of the laboratory was focusing the randomness of how many gene fragments that are sequenced (random within a range of 25-50 million in the case study) and which gene fragments in the sample that are actually sequenced in the sequencing machine (completely random process).

In this study, we assumed that the bioinformatic pipelines to assign the sequenced gene fragments to AMR genes was perfect (which is actually not correct assumption).

2.4 statistically analyses of observed data

The data obtained in each iteration was analysed statistically with the aim to detect changes in the occurrence of AMR (breakpoint analysis) and to forecast the occurrence of AMR beyond the latest time point of sampling.

Initially different techniques for breakpoint analyses and forecasting was explored. A statistically method, originally developed to both detect and forecast changes in vegetation using remote information, was utilised - Detecting Breakpoints and Estimating Segments in Trend (DBEST). In this method, potential change points in the results and change in the amount of resistance of the chosen genes over a 20-year period is identified by performing a time-trend-analyses looking for changes in the measured amount of the AMR-gene(s) of interest over time.
The main idea behind DBEST is to detect changes in a trend component. The trend component of the time series is then segmented using the peak/valley detector function. This method determines by drawing a straight line through detected peak/valley points and comparing perpendicular distances to non-peak and non-valley points between them using a distance-threshold. In this case, we use DBEST's change detection algorithm, which checks if the trend has segments with variation less than the threshold value set for the magnitude of a change and, as a result, identifies a final set of breakpoints with magnitudes greater than this. Data type, seasonality, change detection, first level shift, second level shift, duration, distance threshold, alpha, and change magnitude are the primary parameters used in this change detection algorithm within DBEST (Jamali et al., 2015b, 2015a). The parameters used for DBEST are shown in Table 1.

Table1: The parameters used for DBEST

First-level-shift	*True_Mean/2
Second-level-shift	*True_Mean
Duration	6 months
Alpha	0.5
Change magnitude	10% of *True_Mean

*True_Mean is the true mean (average) of each AMR-gene data used in this work.

In scenario 1, all detected breakpoints during the initial first 120 months is false alarm, given that in the true data there is no change. Breakpoints that were detected after months 121 were used to estimate the time to detection in each of the iterations, representing the time that can be expected between the initial emerge of AMR of interest and the actual detection of the emerge.

2.5 Modelling surveillance programme

The model we have developed gives the opportunity to assess the effect of factors that we can decide to change – number of samples taken and polling, the sampling sequence, how many DNA fragments we want to sequence in a sample, and how we want to analyse and interpret the observed results. This assessment is done in the context of reality taking into account how the gene of interest (in this case study AMR genes) vary between farms, within between animals within farms and a change over time.

Table 2: Different scenarios for genes occurring everywhere in the population (beta lactam and tetracycline)

Scenario	No. of faecal samples per (one pool per month	Duration (months)	Initial occurrence (%)	Increased (%) after 120 months (per month)	Sequence depth (millions)
1	5	240	100	5	25-50
2	20	240	100	5	25-50
3	100	240	100	5	25-50

Table 3: Different scenarios for a newly introduced AMR gene

Scenario	No. of faecal samples per (one pool per month)	Population size (farms)	No. of seed farm	Duration (months)	Sequence depth (millions)	Transmission rate (*SI*) model
1	5	4000	1	60	25-50	0.148
2	20	4000	1	60	25-50min	0.148
3	100	4000	1	60	25-50	0.148

2.6 Sensitivity analysis

To assess the relative impact of each scenario on the spread of the AMR-gene within the farms, sensitivity analysis is required. R software was used to perform all statistical analyses. We incorporate stochasticity into farm population to account for certain levels of unpredictability or randomness, as demonstrated by individual animal movement within the farm. The model predictions are highly robust (as shown in figure 4 and 8), according to sensitivity and robustness analyses performed on model input parameters. The transmission rate of the AMR-gene, as well as the proportion of the sampling schedule, were identified as potentially influential scenarios in the analyses.

3. Results and discussion

In this study, we have measured the occurrence of AMR genes as the counts per million, that should be interpreted that how many of the sequenced gene fragments was representing a specific AMR gene (normalized to 1.000.000). In DNA sequencing many different measures can be calculated and choice a unit depends on the aim of the study. The unit from other DNA based laboratory techniques have other units – e.g. Ct values of RT-PCR. Although the information might be quantitative from many new laboratory technologies, it is important to make careful considerations about the scale of measurement, before performing statistical analysis and interpretation of the results.

Figure 1 shows the results of one iteration (out of totally 101 iterations) with the sample size of 5, 20 and 100, respectively, from the scenario of changes of beta-lactam resistance in the population. The outcome of each iteration (as presented in figure 1) was used to detect the possible break points and false alarms in each iteration. Figure 2 shows the distribution of time to detect in the 101 iterations, which ranging from a detection within the first few months after increased occurrence to that the increase is detected first after several years. This variation in time to detection is a result of all the sources of variability in the surveillance program ranging from sampling schedule to the definition of change in the statistical analysis. Figure 3 shows number of false alarms in the iterations. In most iterations the number of false alarms is 0 or only 1 or 2, but there are also iterations with several false alarms.

Figure 4 shows the forecasted occurrence of beta-lactam AMR based on the observed data in the surveillance. The size of the uncertainty is increased by the time period that is forecasted, but instead of an eye-ball forecast, an analytical approach utilizing the data is more valid, robust and objective.

Figure 5, figure 6, figure 7 and figure 8 present equivalent results for the scenario for changes in tetracycline resistance.

In figure 1 and 4, on top of the simulated data, the output from the statistical analysis performed to detect changes are presented – detected changes and type of detected changes based on the definition of abrupt change.

The DBEST algorithm employs a trend estimation method derived from the Loess Seasonal-Trend decomposition procedure (STL). STL is a fast filtering procedure that can deal with missing values and divides a time series into trend (low frequency variation), seasonal (variation at or near the seasonal frequency), and remainder (remaining variation) components. If one or more level-shift points are detected, the STL decomposition is performed separately for each part of the time series divided by the detected level-shift points; otherwise, it is performed once for the entire time series. The separate STL decomposition procedure frequently produces more precise trend and seasonal components, particularly around detected level-shift points. The observed data was deseasonalized by looking for the trend component, level-shifts are detected in the same way, but no STL decomposing is required.

Based on the studied application, we define what properties a data point must have in order to be considered as an abrupt change or a level-shift. This is accomplished through the use of three arguments: first-level-shift, second-level-shift, and duration. For more information, see (Jamali et al., 2015b, 2015a). The values for these arguments must be careful considered. The presented tool gives the opportunity to assess the effect of these definitions on the number of false alarms and time to detect.

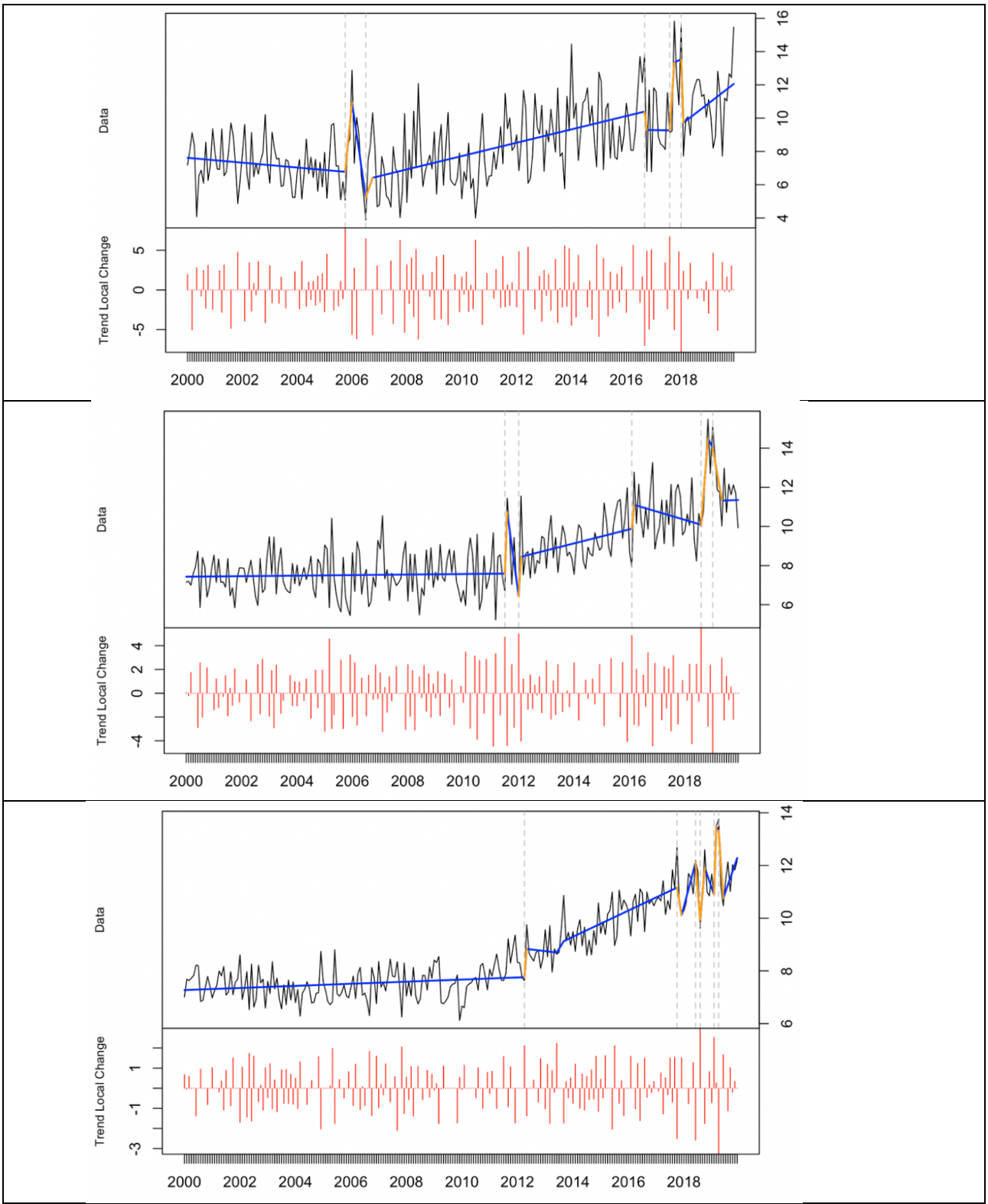


Figure 1: Results from surveillance of increased occurrence of beta lactam resistance. top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool. Black: the observed data points. Blue: the fitted trend. Orange: the gradual changes detected. Dashed vertical lines mark the starting point of detected changes.

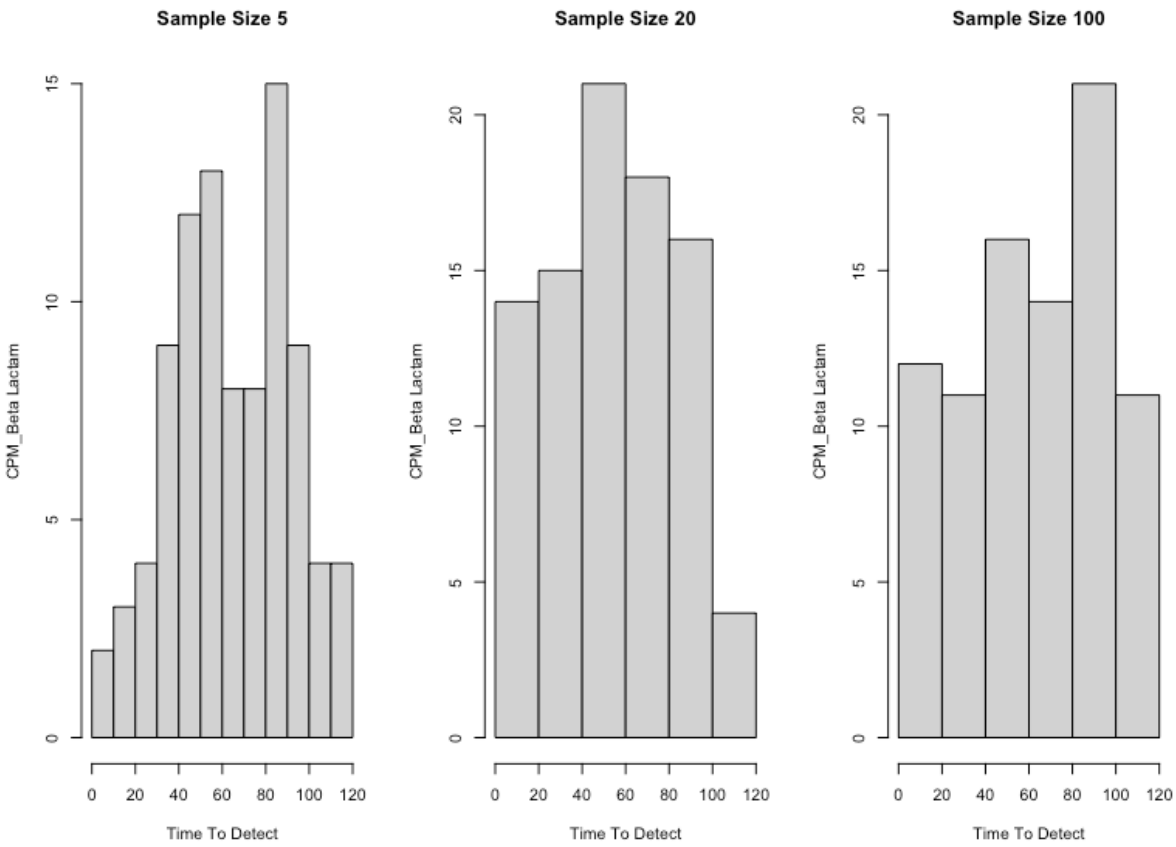


Figure 2: Time to detect changes in the occurrence of beta-lactam [months] in the 101 iteration performed for a sample size of 5 per pool (left), 20 samples per pool (middle) and 100 samples per pool per month (right).

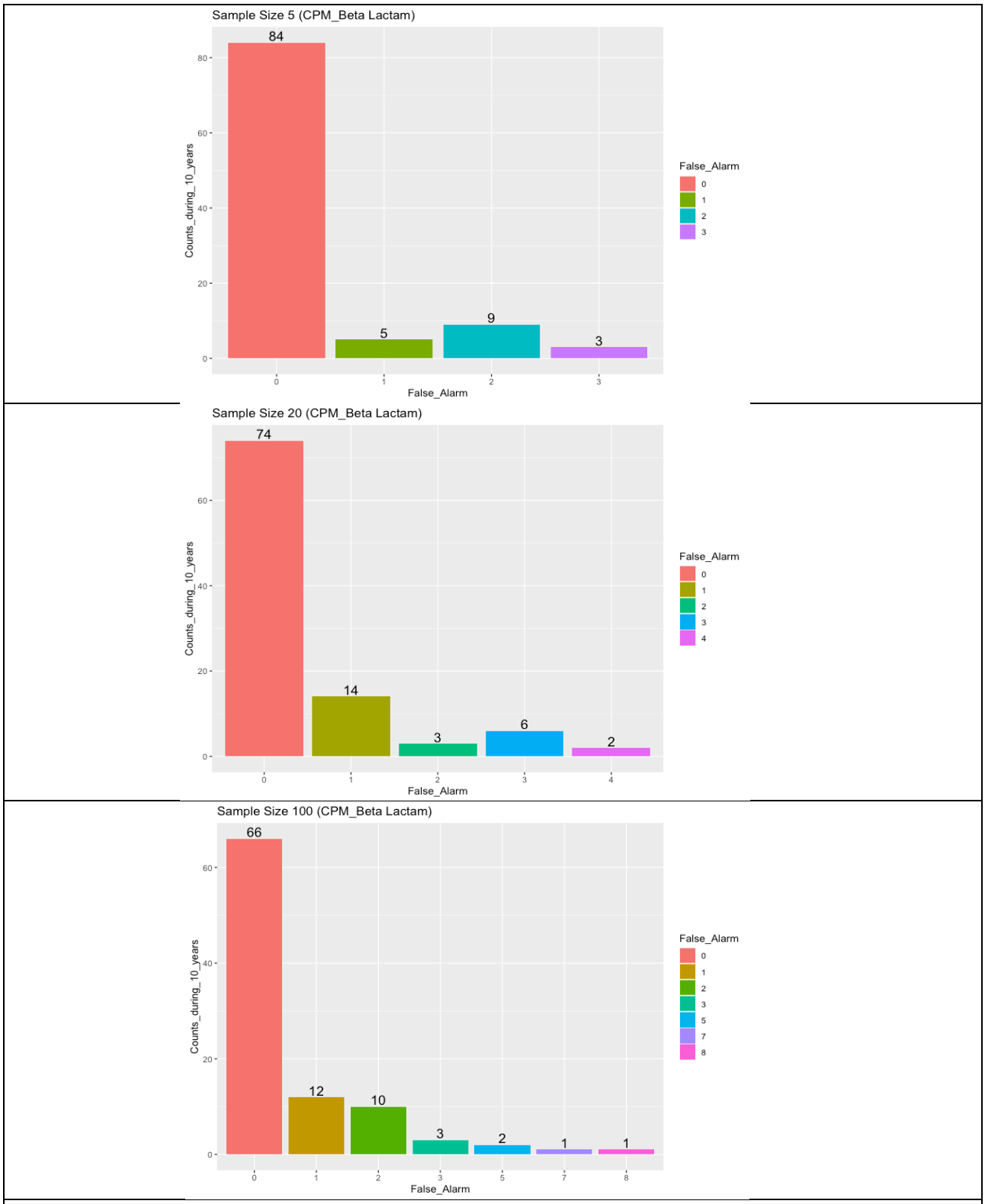


Figure 3: Number of false alarms (detected changes in observed data without changes in the true occurrence) in the surveillance of increased occurrence of beta lactam resistance. Top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool.

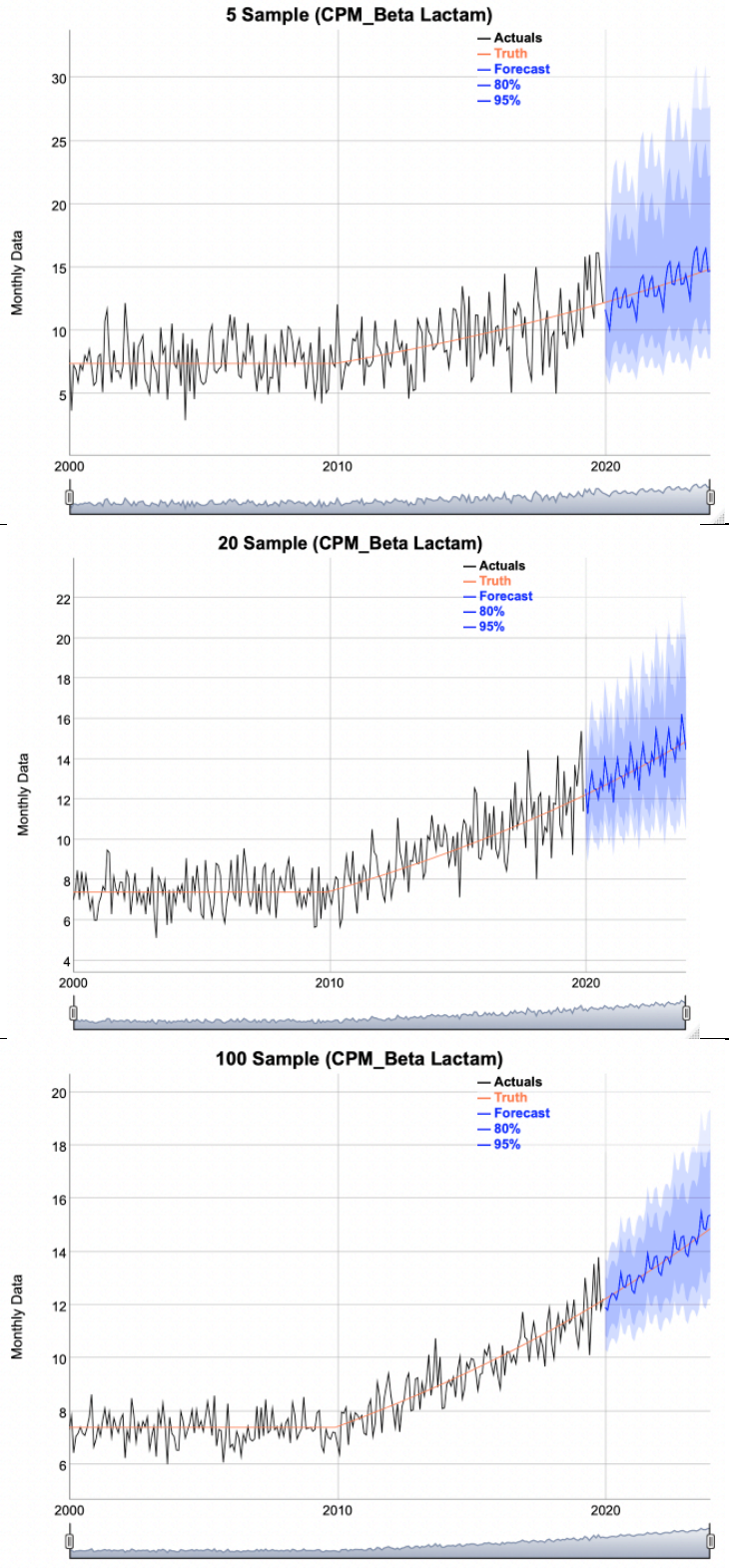


Figure 4 shows the true occurrence of beta-lactam resistance, the observed occurrence forecasted occurrence of beta lactam. The shaded regions around the forecasted case trend (deep blue line) represent the 80% (deep blue) and 95% (light blue) confidence intervals. Top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool.

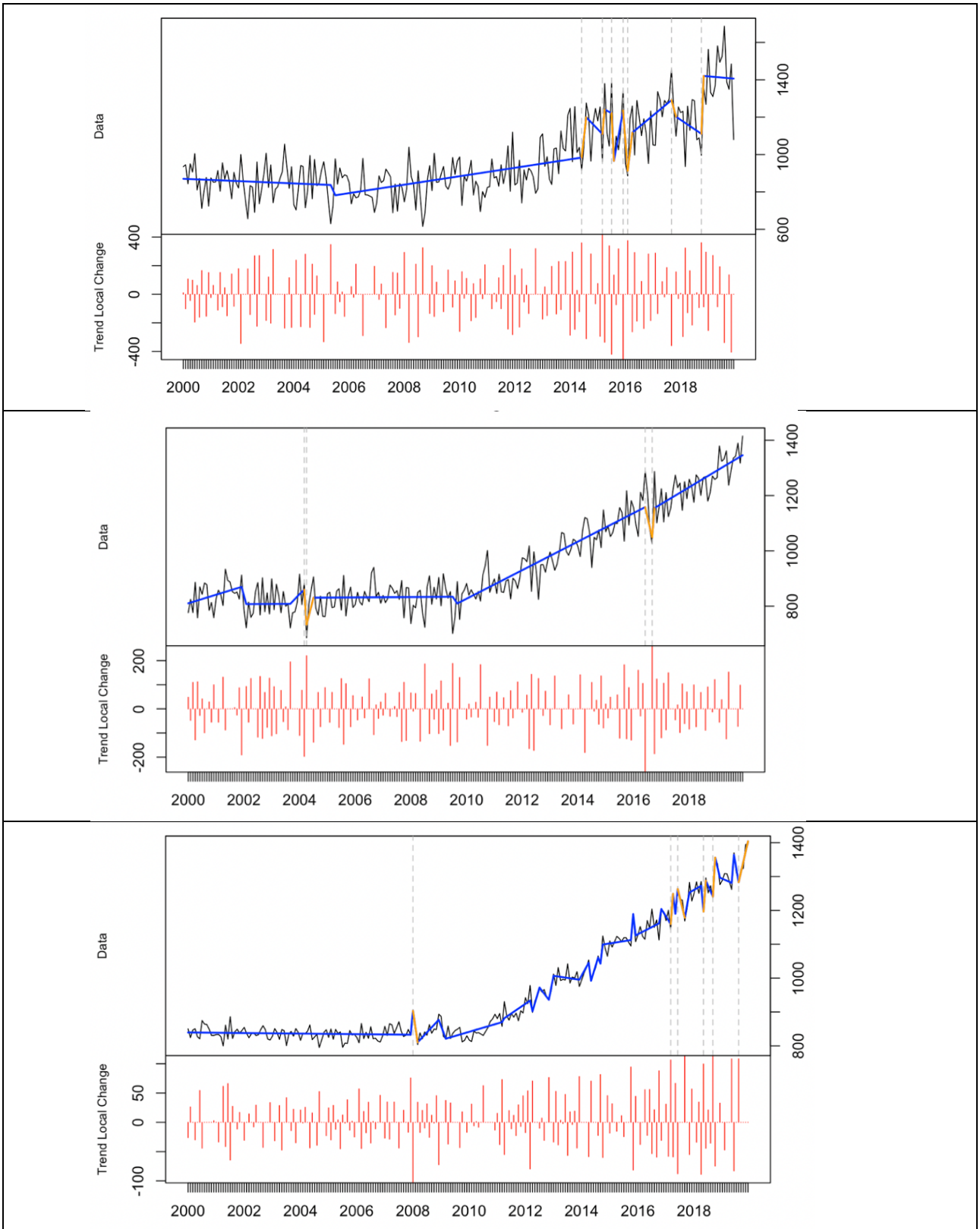


Figure 5: Results from surveillance of increased occurrence of tetracycline resistance. top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool. Black: the observed data points. Blue: the fitted trend. Orange: the gradual changes detected. Dashed vertical lines mark the starting point of detected changes.

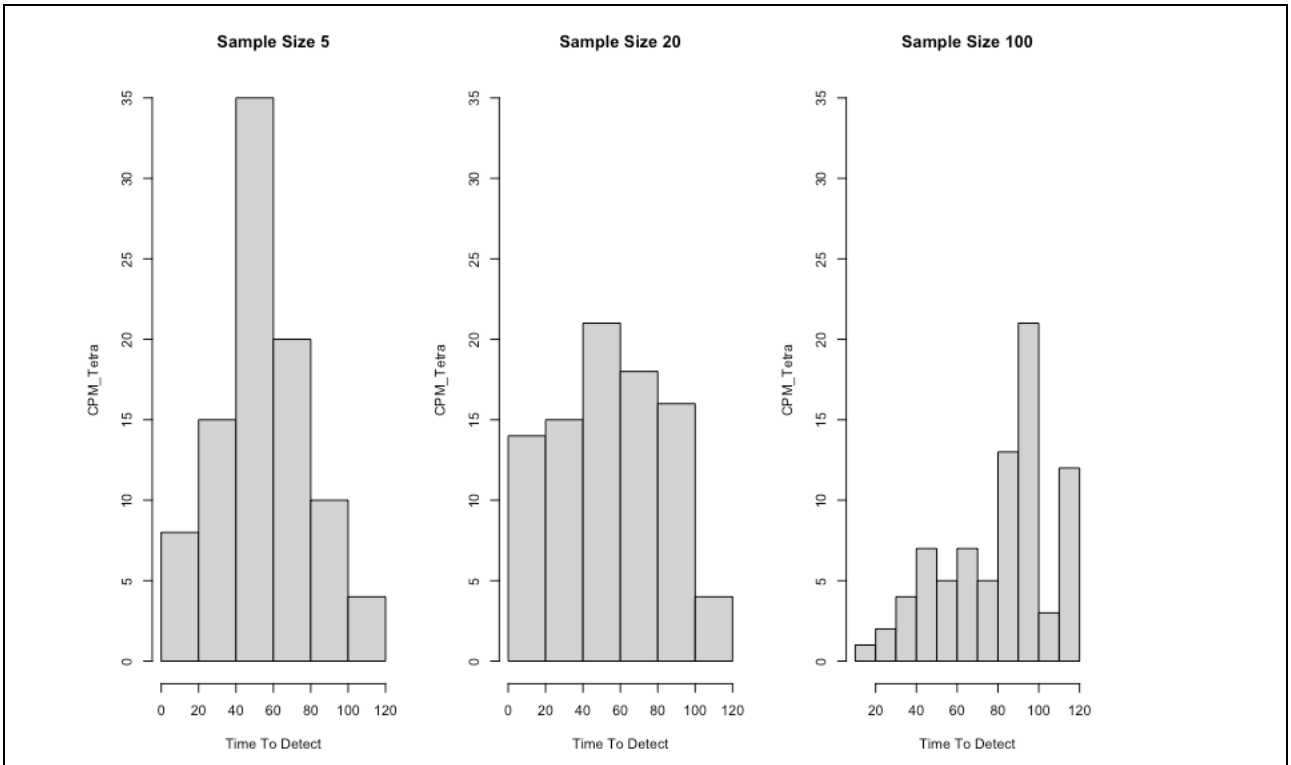


Figure 6: Time to detect changes in the occurrence of tetracycline [months] in the 101 iteration performed for a sample size of 5 per pool (left), 20 samples per pool (middle) and 100 samples per pool per month (right).

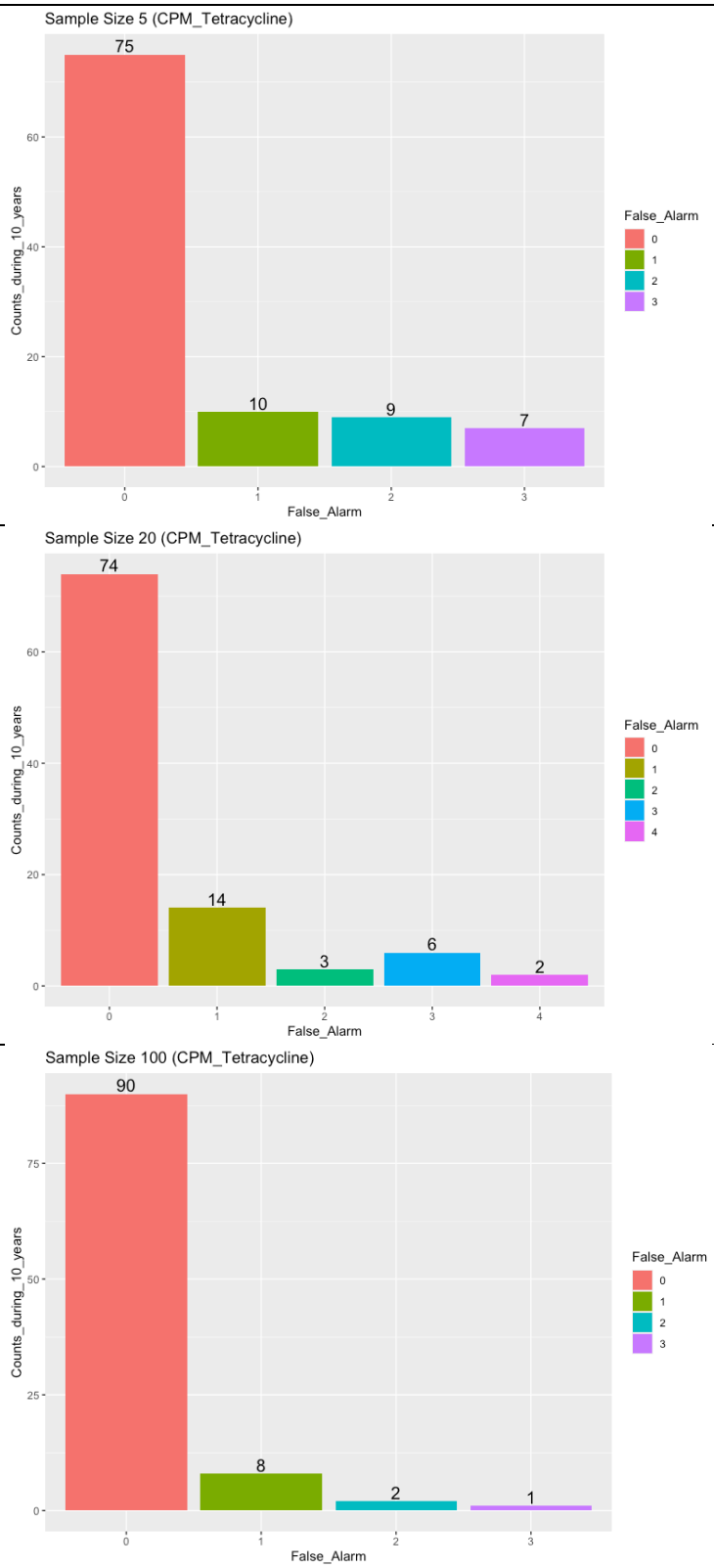
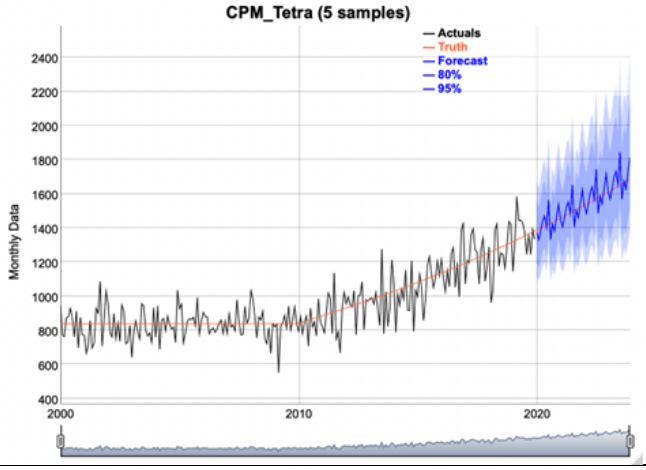
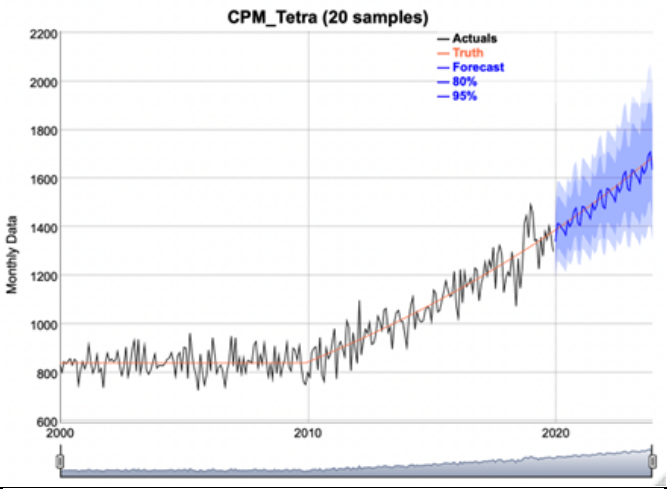


Figure 7: Number of false alarms (detected changes in observed data without changes in the true occurrence) in the surveillance of increased occurrence of tetracycline resistance. Top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool.





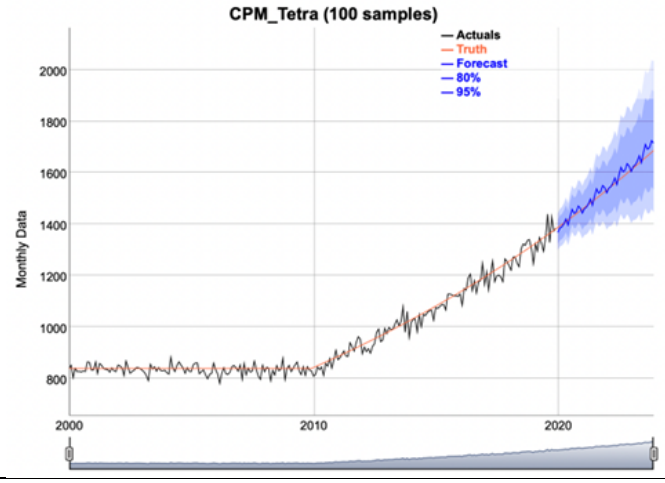


Figure 8 shows the true occurrence of tetracycline resistance, the observed occurrence forecasted occurrence of beta lactam. The shaded regions around the forecasted case trend (deep blue line) represent the 80% (deep blue) and 95% (light blue) confidence intervals. Top – 5 samples per pool, middle 20 samples per pool, bottom 100 samples per pool.

From Figure 4 and figure 8, the observed data and the forecast is close to the “true data”. This suggests that the data overall performance of the surveillance is a valid representation of the truth both and can be used to forecast the occurrence.

According to Figure 9, scenario 1 shows that the 5-sample size, the more likely it is to detect the gene at an early stage of its spread followed by scenario 3 and then scenario 2.

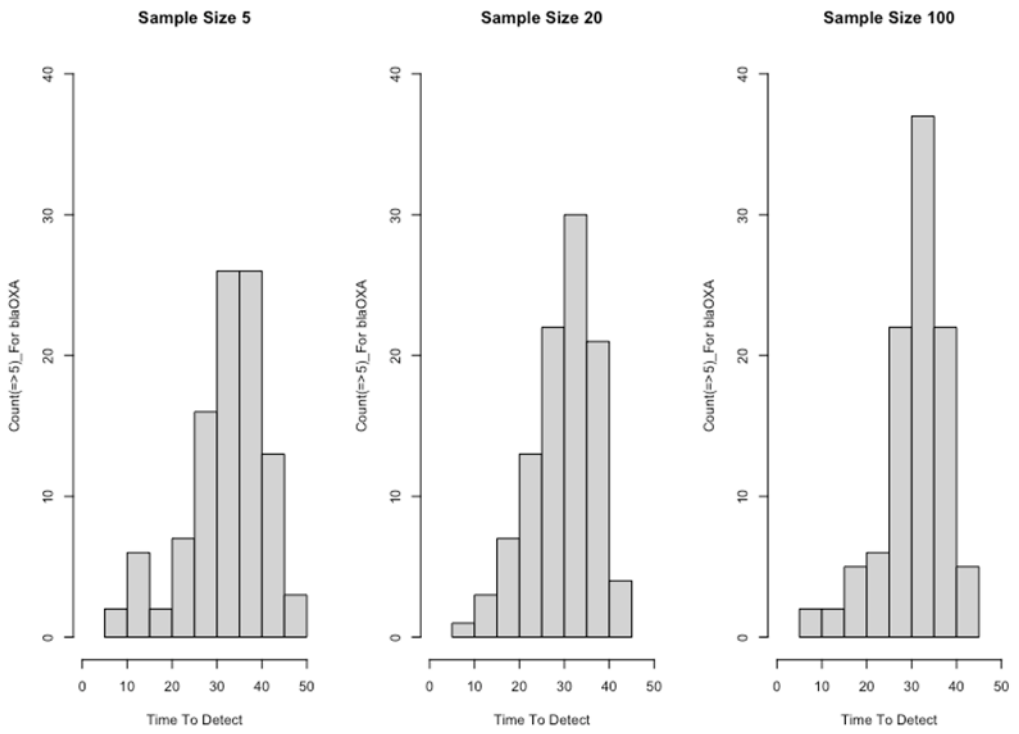


Figure 9: Time to detect a newly introduce AMR gene into the pig population in the 101 iteration performed for a sample size of 5 per pool (left), 20 samples per pool (middle) and 100 samples per pool per month (right). The transmission rate is equivalent to that almost all farms are infected after 5 years. The occurrence of the pathogen within animals in infected animals is assumed to be the same for AMR genes that are present at a very low rate in reality

4 Conclusion based on the modelling work

We have developed a framework for assessing the effect of implementing new laboratory techniques in already existing surveillance programs, and how that will influence the performance of the program. The framework can also be used to optimize the sampling schedule given new laboratory techniques.

The framework presented take into account not only knowledge to the laboratory techniques, but also how the hazard of interest is present in the population of interest, both increase of an existing hazard or an emerge on a new hazard. In the case presented in this report, we work with surveillance of AMR in the animal production and how gene sequencing techniques can be used for that.

The framework takes into account the effect of true sources of variation ranging from randomness in the sampling to random variation in the laboratory techniques.

Also, as a novel part, we also include an analytical method for forecasting in the framework, utilizing computer intensive analytical methods for predicting the future change of the hazard in the population based on trends in observed data. With the computer intensive approach, a more real-time surveillance if achieved, with a continuous update of the forecast for every new observed result. In the framework we utilized only one forecasting methods. The area for forecasting should be investigated further.

Acknowledgement

The authors would like to express their appreciation to Dr. Sadegh Jamali of Lund University's Department of Physical Geography and Ecosystem Science for his constructive feedback. Sadegh will be a part of this project in the future.

Reference

Apenteng, O. O., Arnold, M. E., & Vigre, H. (2020). Using stochastic dynamic modelling to estimate the sensitivity of current and alternative surveillance program of Salmonella in conventional broiler production. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-76514-3

Jamali, S., Jönsson, P., Eklundh, L., Ardö, J., & Seaquist, J. (2015a). Detecting changes in vegetation trends using time series segmentation. *Remote Sensing of Environment*, *156*. https://doi.org/10.1016/j.rse.2014.09.010

Jamali, S., Jönsson, P., Eklundh, L., Ardö, J., & Seaquist, J. (2015b). Remote Sensing of Environment Detecting changes in vegetation trends using time series segmentation. *Remote Sensing of Environment*, *156*.

O'Neill, J. (2016). *Tackling drug-resistant infections globally: final report and recommendations. The review on antimicrobial resistance*. https://amr-review.org/sites/default/files/160525_Final paper_with cover.pdf