Towards an integrated food safety surveillance system: exploring the potential of combining genomic and epidemiological metadata

Hill, A.*, M. Crotta, B. Wall, S. O’Brien, L. Good and J. Guitian

*Contact: andrew.hill9@baesystems.com

Deliverable fulfilling contract SEP-EOI-03

Main supplier:

BAE Systems (Operations) Limited Registered No: 1996687

Executive Summary

Background and aims

Foodborne infection is a result of exposure to complex, dynamic food systems of ever-changing trade routes, production processes, consumption patterns and pathogen evolution. The efficiency of foodborne infection, and the virulence of that infection, will be driven by the ongoing shift in the genetic machinery that power bacterial and viral pathogens. Next Generation Sequencing (NGS) technologies, in particular Whole Genome Sequencing (WGS), provide very high resolution data about the genetics of a pathogen and its capability to infect and cause illness in people.

While the supplementation of genetic data with epidemiological metadata (for example, phenotype, case, geographic data, but more widely any data that allows WGS data to be categorised by the food chain or other environmental context) is encouraged by the European Food Safety Authority (EFSA) and others, there is a distinct lack of data describing the food chain upon which a pathogen can transmit (or has transmitted). Indeed, EFSA themselves note that “at present, prototype databases cannot be used for surveillance purposes since they are not widely linked to epidemiological data”. Put simply, the high resolution of NGS technologies is not matched by the complementary metadata.

The original project aims were i) to review the state-of-the-art in NGS and big data technologies, ii) to produce a prototype framework within which epidemiological metadata could be collected with genetic data to improve the efficiency of food safety surveillance and control, and iii) to start to identify relevant “big data” data and methods that could potentially be used to conduct the necessary analyses. These aims were completed, but were achieved in a slightly modified fashion than originally envisaged. As our thinking developed throughout the project, we developed a working hypothesis that to derive genuine knowledge about risk we must accompany high-fidelity genetic data with similarly high-fidelity metadata. We therefore decided to concentrate on developing a hypothetical food chain model (as we had no access to genetic or relevant metadata) and to identify the impact of progressively better metadata.

Predictive analytics

The great promise of the explosion in big data is that we can start to better inform decision-makers by interrogating large (unrelated) databases for trends and patterns that can then be used to forecast what will happen in near real-time (see figure below). From a food safety view we largely use human case data to record what has happened (descriptive analytics). Early detection algorithms and epidemiological investigations allow us some insight into what is happening (diagnostic analytics), but we are a long way from being able to predict what will happen (predictive analytics), or how we can prevent an impending outbreak.

The big difference between predictive algorithms derived from big data and classical horizon scanning is the focus on identifying patterns in amongst huge datasets, and moving from a more qualitative output to a quantitative prediction. In our judgement, we are arguably some years away from truly having foresight or preventive capability, but the technologies to provide the big data required do already exist – it is a matter of understanding how best to
use these big data, and implementing an efficient infrastructure to generate the best, most relevant metadata.

Project analysis

Within the timescale of the project we were unable to access any genetic data sources. We therefore concentrated on building a hypothetical food chain, built so that we could keep track (with absolute precision) of the entire chain of information from contamination through processing to infection, including 100 varying genotypes of a pathogen X. The output of this model was a data array that recorded all of the parameter values (e.g. temperature of processing, the burden of contamination of each of the 100 genotypes) and whether a processed food unit led to human infection. That is, we had complete knowledge of the hypothetical food chain and how variations in that food chain influenced human infection. We then used various parts of this complete dataset to test the accuracy of progressively more sophisticated food surveillance systems, starting with a system similar to the current process, and ending up with a very data-rich system where many if not all food servings could be traced back along the food chain.

The lack of contextual data meant that current surveillance system activities such as early detection algorithms are relatively insensitive. Current epidemiological data gained from outbreaks (such as they are) do not readily lend themselves to robust analysis that can then be used to improve wider scientific knowledge. The inclusion of WGS information alone within current microbiological sampling frameworks at processing did not advance an ability to correlate contamination with cases. Only the inclusion of efficient tracking and tracing capabilities allowed the successful prediction of higher-risk interactions of environment and genetics. Foodborne infections are relatively rare within the context of the vast numbers of meals consumed – in our model there was a ratio of roughly 1 reported case of infection per
1000 units consumed. This means that a lot of data need to be collected to have a robust dataset of cases from which to analyse.

Big data technologies, for example RFID tracking and networked sensors, allowing tracking and tracing along the food chain as well as allying metadata to products, would provide a way of enabling the collation of big data along the food chain. Machine learning methods, such as the decision tree algorithm applied in this project, could then be used to identify patterns and trends in data. The prototype analysis conducted during this project suggested such that a networked food chain feeding metadata into a machine learning algorithm would have reasonable sensitivity and specificity for predicting cases. The main aim of machine learning is to produce predictive models, but at this early stage of research it is probably safer to suggest that these methods would allow us to analyse the combined genomic/meta data to identify dangerous interactions of environmental and genetic characteristics. This information could be used for both long-term research informing food safety policies, and to provide contextual data to supplement current early detection algorithms. A potential framework for the inclusion of high-fidelity NGS (i.e. WGS) and metadata into a surveillance framework is highlighted in the figure below. KPI1 and KPI2 are top-level key performance indicators.
Conclusions

1. As with any food chain, the burden of contamination on food is generally more important than the specific genetic makeup of the pathogen. The relative contribution of genetics to risk is in most cases probably less than the contribution of the food chain and the variation in processing practices and human behaviours. Therefore, we must consider the interdependencies between the genetic and physical epidemiology of foodborne infection if we are to have a true understanding of risk.

2. Given this large random “noise” in complex food chains due to physical and environmental variance, it is difficult to tease out useful insights from WGS data alone. Even having WGS information at both processing and infection stages, within a very simple hypothetical food chain (real food chains being subject to far more complexity/randomness), did not allow much further insight to be generated. Therefore developing source attribution models based on WGS and pathogen data alone is subject to large uncertainties.

3. Adding physical characteristics of the food chain (e.g. temperature control) is unlikely to be beneficial unless there is much better tracking of animals/ingredients through the food chain. Where we were able to match with precision each food unit to an infected consumer, we were able to start to identify plausible relationships and interdependencies between phenotype and genotype.

4. However, even with our simple food chain, it was not possible to produce an acceptable predictive model. This is a cautionary note: the rarity of foodborne infection means that even predictive models with high sensitivity and specificity have low predictive power due to a large number of false positives.

5. The benefits of combining epidemiological big data with genomic data should be in recognising the relative contributions of the physical food chain together with genomic data. Machine learning approaches could be applied to understand higher-risk profiles – matching higher-risk genes to problematic processing conditions – which could then be used to inform horizon scanning activities.

Benefits

The specific benefits of this preliminary work are:

1. We reviewed the literature on both genomic and epidemiological metadata (original Aim 1 as described above). Much of the work required has already been conducted by EFSA, as described in recent scientific opinions on the subject of WGS, which states that appropriate epidemiological metadata must be attached to WGS data to be usable in surveillance and control contexts. We furthered this work by investigating the type of epidemiological metadata that would be required, and came
to the conclusion that better tracking and tracing is a pre-requisite for improving epidemiological metadata.

2. We developed a hypothetical food chain model that allowed us to interrogate various surveillance frameworks without the need for real genetic data. In addition, this provided us with a completely known system, so that we could validate the results of the various analyses. This model can be expanded to represent other, more complex food chains for future, more in-depth analyses.

3. We showed that high-fidelity metadata are likely to be needed if high-fidelity NGS or WGS data is to be used to develop a genuine understanding of overall risk and the interactions of environment and genetics.

4. By stepping through various potential datasets as surveillance sources, we showed the advantages and disadvantages of different data sources. We therefore now have a clearer idea of the metadata required to improve food safety surveillance (Aim 3).

5. By applying machine learning methods to our dummy data, we have shown the relevance and capability of “big data” methods, which can be used to combine genomic and metadata and identify dangerous interactions.

6. We have suggested a top-level framework for how metadata could be combined into the current surveillance systems (Aim 2).

Wider benefits of combining metadata with genomic data are:

1. Getting the right metadata not only improves our long-term understanding, but also maximises the value of NGS technologies and the investments required in them.

2. Actively monitoring the food chain could move us from reporting the past (weekly summaries of cases) to forecasting high-risk scenarios (predictive models), or at the very least supplementing current surveillance systems with better information for epidemiological investigations.

3. Food chain metadata could allow us to react quicker to outbreaks (by better and rapid tracking and tracing).

Recommendations

Our recommendations are split into short, medium and long-term activities.

Immediate

1. Meet with Warwick PL to discuss potential synergies and attend their workshop to present results;

2. Through collaboration with Warwick, produce a combined set of benefits and recommendations from both projects;

3. Refining our working hypothesis to better test it – for example can we specify in greater detail what level of track and trace resolution is required (e.g. processing plant, batch, food unit)?

4. Further investigation of proof-of-concept machine learning techniques, for example:
   a. Conducting machine learning analysis on a more complicated hypothetical model;
   b. Other predictive analytic methods, such as neural network analysis, which could be used to identify risky combinations of genotypes and production factors.
5. Identify 3-4 areas of research where combination of real WGS data (perhaps from GMI) and big data would be beneficial.

Medium-term

1. Development of Performance Indicators for food safety surveillance (FSA strategy objective);
2. Further investigation of relevant metadata (for GMI) to enable measurement of Performance Indicators;
3. Develop infrastructure required to measure and assess Performance Indicators;
4. Discussion with industry over use of WGS and “big data” technologies, including track and trace, generation of networked metadata, and any access issues.

Long-term

1. Implementation of new logistical technologies into food production chain;
2. Agreement over right level of discrimination for use of WGS in surveillance systems and associated research;
3. Central data collection of metadata into GMI or a compatible database;
4. Implementation of new surveillance algorithms using both case data and epidemiological data.
# Contents

1 Introduction ................................................................................................................................. 1

2 Overview of methodology ............................................................................................................ 3
   2.1 Objectives of food safety surveillance system ....................................................................... 3
   2.2 Predictive analytics ................................................................................................................. 4
   2.3 Analysis framework .................................................................................................................. 6

3 Generating baseline simulated dataset of human infections and related food unit parameters ................................................................................................................................. 7
   3.1 Choice of hypothetical food chain model .............................................................................. 8
   3.2 Mathematical model ................................................................................................................ 9
      3.2.1 Modification and extension of Swart processing model .................................................. 9
      3.2.2 Genotype generation ....................................................................................................... 12
      3.2.3 Output of the hypothetical food chain model .................................................................. 16
      3.2.4 Parameter estimation ..................................................................................................... 17
      3.2.5 Implementation ............................................................................................................... 17

4 Surveillance scenario analysis ...................................................................................................... 20
   4.1 Hypothesis and data-driven exploratory analyses ................................................................. 20
   4.2 Scenario 1: WGS data alone – early detection algorithm ...................................................... 20
   4.3 Scenario 2: WGS data alone - WGS collection during microbiological sampling at processing ............................................................................................................. 21
   4.4 Scenario 3: Classical statistical analysis of epidemiological metadata and genomic data .......................................................................................................................... 22
   4.5 “Big” epidemiological data - track and trace individual food units ...................................... 23

5 Results .......................................................................................................................................... 25
   5.1 Exploration of simulated data array ..................................................................................... 25
   5.2 Scenario analysis ................................................................................................................... 29
      5.2.1 Scenario 1: WGS data alone – early detection algorithm ............................................... 29
      5.2.2 Scenario 2: WGS data alone – addition to microbiological sampling criteria ............... 32
      5.2.3 Scenario 3: Classical statistical analysis of both epidemiological and genomic data .... 33
      5.2.4 Scenario 4: “Big” epidemiological data - track and trace individual food units........... 35

6 Discussion .................................................................................................................................... 38
   6.1 Discussion of results ............................................................................................................... 38
   6.2 How “big” epidemiological metadata could be used to supplement current food safety surveillance .................................................................................................................. 41
   6.3 Implementation (barriers and practicalities) ......................................................................... 42
   6.4 Next steps and recommendations for further research .......................................................... 44
6.5 Recommendations........................................................................................................44
7 References .....................................................................................................................46
8 Glossary and abbreviations..........................................................................................49
1 Introduction

The most important enteric disease burdens worldwide are caused by microbiological hazards such as *Salmonella spp.* and *Campylobacter spp.*, which can cause human infection via both food and environmental routes. In the United Kingdom (UK), good progress has been made in reducing the number of Salmonella cases over the past couple of decades, although Campylobacter cases remain stubbornly high \([1,2]\). The large burden of disease reflects the exposure of people to a complex food chain of ever-changing trade routes, production processes and consumption patterns that interact with the ongoing evolution of enteric pathogens. The efficiency of zoonotic infection, and the virulence of that infection, will be driven by the constant shift in the genetic machinery that power bacterial and viral pathogens. Therefore, both the food chain and genotype should be considered when assessing zoonotic risk. To date, research into the combination of food chain data with genetic pathogen data has been scarce; indeed most papers referencing genetic epidemiology simply state the case for the use of genetic information in epidemiology, rather than suggesting how to combine data sources \([3–5]\). In this paper, we present a hypothetical, quantitative case study, focusing specifically on meatborne infection, to highlight the value of combining both genotypic and epidemiological metadata.

Traditional identification or typing systems based on phenotype (i.e. serotype, biotype, phage-type or antibiogram), have been used for many years. Since the initial use of plasmid fingerprinting in the 1980s \([6]\), the molecular characterisation of bacterial isolates revolutionised the ability to differentiate bacterial subtypes and has become an essential component for epidemiological investigation of infectious disease, enhanced surveillance and detection and source attribution of outbreaks caused by foodborne pathogens \([7,8]\). Whole Genome Sequencing (WGS) is part of a revolution involving so-called Next Generation Sequencing (NGS) technologies, a set of methods that can reveal the entire genome of an organism. In-field applications of WGS have garnered success in multiple contexts in recent years: from transmission studies and outbreak tracking of pathogens in hospitals \([9–11]\) to traceback investigations and source-attribution in foodborne outbreaks at both national and multinational levels \([13–15]\).

In a recent review by the European Food Safety Authority (EFSA) \([16,17]\) it was recognised that NGS and WGS technologies are not currently used to their full potential with regards to risk assessment and epidemiological insight. The reasons for this include, but are not limited to: i) the construction of the necessary infrastructure to deal with the volume of sequencing data generated, ii) deciding which metrics are useful in evaluating risk, and iii) what
metadata should be collected with the genomic data to set the context of the system in which
the WGS sample was taken. Each of these reasons alone represents a formidable challenge
in turning genomic data into actionable knowledge that can be used to improve surveillance
and control. EFSA notes that ‘at present, prototype databases cannot be used for
surveillance purposes since they are not widely linked to epidemiological data’. Put simply,
the high fidelity and resolution of NGS technologies is not matched by the complementary
metadata.

In parallel to the explosion in available genetic sequencing data, another revolutionary
technology is the networking of systems and sensors and the generation of big data
(colloquially known as the internet of things). However, this revolution in “big data” is largely
unexplored in food safety research. Networked sensors can provide real-time, centrally-
collated data on the performance of production systems as is used in various industries,
from energy to car manufacturing. In addition, Radio Frequency Identification (RFID) tagging
provides the ability to track and trace individual items, as well as providing a wealth of data
for individual products. The application of these technologies could potentially revolutionise
food production efficiency and quality, and indeed several trials of tracking and tracing using
RFID have been published \[^{18-20}\]. These technologies provide the potential to generate a
similarly high-discriminatory power for epidemiological data as for NGS technologies, which
we believe is a key missing link in translating WGS data into actionable knowledge.

We must strike a cautionary note when discussing the potential of big data: it is sometimes
promoted as a panacea to almost all analytical problems, but we must always be mindful of
the need to construct decision support evidence (using big data or otherwise) with analytical
and statistical rigour. Just because there is a large dataset to work with does not mean it is
representative of the study population or contains all relevant contextual metadata. Our
working hypothesis is therefore that we must have high-fidelity but relevant epidemiological
metadata if we are to genuinely transform genomic data into actionable knowledge that can
assist policy makers to make informed decisions about food safety risk. There will still be
limits on the insight that any feasible surveillance system can generate, especially as most
food chains are complex and subject to a large volume of noise in the system. Therefore, our
aim in this paper is to begin to explore our hypothesis by using a simple food chain model to
investigate how detailed epidemiological data may fit into a food safety surveillance system,
and ultimately to begin to assess the potential benefits that such data may bring for public
health risk assessment.
2 Overview of methodology

2.1 Objectives of food safety surveillance system

WGS and NGS technologies provide unprecedented discriminatory detail, but there remains a question over how best to apply this knowledge within a risk-based surveillance system that should primarily be geared to preventing human exposure to large doses in food items. Dividing up total burdens of a pathogen into ever more discrete subtypes may increase the sensitivity of detecting outbreaks, but alone it is unlikely to be especially helpful for a general surveillance framework that aims to reduce overall foodborne illness. However, being able to identify, evaluate and prioritise novel and higher-risk pathogens/genes within the diverse range of foodborne organisms would be a key contributing factor to mitigate along the food chain and to prevent the (re-)emergence of potentially dangerous pathogens.

Broadly speaking, there are two main objectives of a food safety surveillance system that monitors the incidence of foodborne disease: i) to inform policies that reduce the number of cases over time and ii) to identify and control foodborne outbreaks as and when they occur. These two objectives are not mutually exclusive, but differentiate the emergency response and the longer-term ongoing refinement of any national surveillance system.

These two objectives may be summarised as top-level KPIs thus:

KPI 1: Reduction in outbreak-related foodborne illness (via improvement of the percentage of cases that can be traced to source);

KPI 2: Annual reduction target in cases of foodborne illness.

The specific KPIs, to reduce the outbreak-attributable (estimated) number of cases, and the overall burden of foodborne illness, should be the ultimate measure of success for a food safety (surveillance) system, which may also have sub-categories such as reducing overall morbidity or mortality. However, certainly in the case of KPI1 (outbreak-related cases), maintaining a good estimate of prevented cases would be a difficult challenge that is inherently uncertain. Other supporting metrics, such as the sensitivity of the surveillance system, the timeliness of outbreak detection, and the proportion of suspected foodborne outbreaks for which a definitive diagnosis is reached, would likely prove more amenable to measurement and provide a reasonable track record of progress in the short-to-medium term.
For KPI 2, we would expect a reduction in laboratory-confirmed cases to be a reasonable measure of success in reducing overall foodborne illness. There are various activities that contribute to the generation of knowledge to improve the efficiency of food safety measures across the food chain, from farm inspections, meat inspection, inspection of catering establishments, through to industry quality assurance schemes. One key problem to resolve is that while the aims of these activities are all targeted towards producing safer food, there is a general lack of co-ordination across government and industry in collecting the “right” data and implementing co-ordinated data strategies across the food chain. Without this it is difficult for researchers and analysts to generate the evidence required by policy makers to develop genuinely successful policies that reduce foodborne illness. Therefore, one of our primary objectives is to start to test what epidemiological metadata actually adds extra value to the high-resolution WGS data, so that FSA may start to co-ordinate its collection.

2.2 Predictive analytics

The advances in big data generation within the past decade or so means that it is now possible for food safety scientists and analysts to quickly, and relatively cheaply, analyse huge datasets. The great promise of this increase in analytical capability is that we can start to better inform decision-makers by interrogating large databases from disparate sources for trends and patterns that can then be used to forecast what will happen in near real-time. We modify a classic description of predictive analytics [21] for food safety in Figure 1. The basic principle is that with big data, we have enough data to start to identify trends and patterns in the data without having to first specify the model that describes how inputs translate into outputs. The theory being that we have enough data that the signals can be found despite the inevitable noise in any relatively complex system. There is also an assumption that huge datasets essentially represent the true underlying distribution of variance in the system, whatever that system may be. These two assumptions may or may not be invalidated depending on the quality of the data.

From a food safety view, if we had big data we could move, in theory, from simply reporting the past (weekly or monthly summaries of how many cases there have been), to describing in more detail what is happening - arguably where we are with the current suite of early detection algorithms - and eventually into the realm of being able to forecast what might happen (predictive analytics) or even to prevent an outbreak occurring (prescriptive analytics). Simply put, such analytical capability could in theory enable us to move from reacting to outbreaks already underway, to identifying higher-risk scenarios and mitigating
against them ahead of time. The big difference between this and classical horizon scanning is the focus on identifying patterns in amongst huge datasets, and moving from a more qualitative output to a quantitative prediction.

Again, we must strike a cautionary note. An obvious tension between the promise and reality of predictive analytics is that these large, potentially unrelated, databases still have to be relevant to the decision in hand. Simply combining available, disparate epidemiological data sources with genetic data, and/or without specific risk questions in mind, is unlikely to result in systematic advances in knowledge. Our main interest therefore lies in understanding the decision-support benefit of combining relevant genetic sequencing datasets with relevant “big” epidemiological data. Without this linkage we are missing the detailed contextual information that allows us to assess risk at even a basic level, yet alone to being able to forecast with any confidence. We are at the beginning of truly acquiring such capability, but arguably the technologies to provide the big data required do already exist: it is a matter of understanding which big data will provide the most added value, and how to generate and use these datasets.

![Figure 1: Progression of knowledge informing the decision-making process with big data and predictive analytics](adapted from Davenport [21]). Within most current food safety surveillance systems, we are currently at the point of reporting the past, with some additional intelligence provided by early warning algorithms and genomic analysis. We are far from being able to make use of the mass of genomic data and other big data to either predict or prescribe events.
2.3 Analysis framework

This preliminary study was aimed at understanding the extra value that can be gained by combining genetic data with epidemiological metadata. All food chains, but especially for meat, are inherently long, complex and dynamic. A serving of meat on a plate has involved the growing, transportation and processing of feed ingredients, the breeding, rearing, slaughter and processing of meat animals, to the distribution, storage and preparation of the meat to consume. Such a long food chain contains many and disparate ingredient sources and processes, with multiple opportunities for contamination of the chain with pathogens and for human/system error to fail to reduce any pathogen contamination to an acceptable level. It is therefore extremely difficult to collect sufficiently robust, routine surveillance data that can be used to separate the "noise" of day-to-day production from the "signal" of malfunctions or gross contamination events that lead to foodborne infection or wider outbreaks. For example, the most detailed information on a foodborne outbreak is likely to come from the sampling and questionnaires of human cases once an outbreak has been identified, potentially long after any evidence for the cause of the outbreak has disappeared or been overwhelmed by noise. Therefore we rarely have a complete understanding of the cause of an outbreak or ongoing infections, and hence it is also extremely difficult to decide what available data would have been useful in source attribution (or what other data could have been useful if they were available).

We therefore developed a hypothetical food chain simulation, so that we can know, with absolute precision, i) the microbiological status of each product throughout the food chain, ii) the performance of the equipment being used at the time of the processing of the product and iii) whether or not it caused a human infection. In this way, we can compare the performance of a surveillance system (with varying levels of incomplete knowledge) against the "gold standard" of our hypothetical, but completely known, simulation. We then investigated what observations we could make about the simulated baseline dataset by assuming different surveillance systems, each reliant on progressively more detailed epidemiological data. The advantage of this hypothetical approach is that we are able to explore the capabilities and limitations of the methodology, and what data are most beneficial to collect, rather than be constrained by the scarcity and quality of the data actually available.

The chronology of the analytical framework was to generate a simulated data array from which to test progressively more advanced surveillance systems (see Figure 2).
The main steps were to:

1. Generate a simulated dataset from the hypothetical food chain model. The unit of interest is a food unit (unit=serving). The output of this model is whether the food unit results in human infection from one or more of G genotypes.

2. Visualise the simulated dataset and understand the relationships between individual parameters and the output. Use these relationships to guide understanding of the outputs of the more advanced analyses.

3. Assume various surveillance strategies based on access to progressively more WGS and/or epidemiological data. Analyse subset of simulated baseline dataset to observe insights that can be generated through collection of various data sources. For example, only matching of an individual food unit to a reported human case would allow access to all the columns of the simulated dataset.

**Figure 2:** Chronology of the analytical framework. A simulated data array, containing relevant epidemiological and genomic data, was generated from a simple food chain model. We then assumed various rows and columns of the data array could be observed with progressively more advanced surveillance systems. We then interrogated the subsets of the data array (representing different surveillance strategies) to determine the relationships and trends observable by first accessing detailed WGS data alone, and then adding the ability to collect detailed epidemiological data via tracking and tracing systems.

3 Generating baseline simulated dataset of human infections and related food unit parameters

Food chains are inherently long, complex and dynamic, meaning it is difficult to collect sufficiently robust surveillance data. For example, the most detailed information on a
foodborne outbreak is likely to come from the sampling and questionnaires of human cases once the outbreak has been identified, potentially long after any evidence for the cause of the outbreak has disappeared or been overwhelmed by noise. Therefore it is rare that we have a complete understanding of the cause of an outbreak or ongoing infections, and hence it is also extremely difficult to decide what available data would have been useful in source attribution (or what other data could have been useful if it was available). We have therefore developed a hypothetical model to test at what points along a specific part of the food chain more data would be useful in understanding the dynamics of foodborne infection and/or outbreaks.

The advantage of a reasonably simple hypothetical food chain model is that we can know, with absolute precision, the microbiological status of each product throughout the food chain, and the performance of the equipment being used at the time of the processing of the product. In this way, we can compare the performance of a surveillance system (with varying levels of incomplete knowledge) against the “gold standard” of our hypothetical, but completely known, model. The hypothetical model is an extreme simplification of reality because i) we only want to show proof of principle and ii) it should also be easy enough to intuitively identify the important trends, hence providing an easy way to validate the effectiveness of our predictive models.

3.1 Choice of hypothetical food chain model

When choosing a model we must first note that the conclusions we are trying to draw are not specific to the activities modelled, but are more general about the extra capability afforded to decision-making by combining genomic and epidemiological data. The same principles of data characterisation and combining genomic and epidemiological data apply whether we are looking at the farm, processing or consumption or any other part of the food chain. We therefore chose a model that allowed us to draw insights around what type of epidemiological data are useful, and the limitations that exist when combining these data with genomic data, but which was also convenient enough to use within this short, exploratory project.

An amenable food chain model should be one that describes in sufficient complexity the performance of industry equipment as well as the inactivation/cross-contamination of zoonotic pathogens (such that we can model the inclusion of both WGS and epidemiological big data). A model with these characteristics was previously developed for an EFSA risk assessment of Salmonella in pigs [22]. In this risk assessment, pigs and their carcasses/food products were modelled individually across the entire production chain, with a high level of
focus on the performance of processing steps, for example scalding and the evisceration process of the pig carcass [23].

We chose two specific stages of the model, scalding and evisceration, as being representative of common food processing practices [23]. Scalding is an example of a hot water bath used in various scenarios during food manufacture, for example to soften hair/feathers on animal carcasses before they are removed, or to partially pasteurise a product. This process is a classic example of both thermal inactivation and washing of a product to produce a critical control point (CCP). On the other hand, inappropriately conducted, scalding is a hazard that can increase the risk of cross-contamination. The evisceration of a carcass can be a gross contamination point if precautions are not taken to minimise the leakage of faecal material from the viscera to the carcass. We generalised these two processes to represent a variety of processing conditions that a food safety surveillance system could potentially capture. We swapped the order of the two processes around, so that we have a gross contamination event first, followed by a CCP stage. For simplicity, we assume that the unit being processed is a typical serving of meat, and the dose received by the consumer is the contamination present at the end of the second processing step. We outline the overall model framework in Figure 3.

3.2 Mathematical model

3.2.1 Modification and extension of Swart processing model

We modified a model developed by Swart et al (2016) [23]. Swart et al do not differentiate bacterial strains from one another, whereas we explicitly modelled one hundred hypothetical genotypes of a hypothetical pathogen X. We therefore reduced the scope of the processing model to two stages so that we could add this additional complexity around the strain variation.

We add a subscript \( g \) to the original Swart equation parameters to incorporate different genotypes (with varying abilities to survive the food chain and cause infection). For simplicity, we assume the individual populations of genotypes act independently from each other. We outline the overall model framework in Figure 3.

*Cross-contamination process (unit \( \rightarrow \) equipment \( \rightarrow \) other units)*

The important aspect of this process is the extrusion of contaminated material from unit \( k \) to the equipment, which then cross-contaminates other units with pathogen \( X_g \) (akin to
evisceration or dehairing). Using coupled recursion equations we can calculate the number of pathogen $X_g$ on the equipment during processing of unit $k$,

$$W_{k+1,g}^1 = \left[ \beta_{1,g} (N_{0,k+1,g} + B_{1,k,g}) + W_{k,g}^1 \right]e^{-\alpha_1 T_{1,k}},$$

$$N_{1,k,g} = W_{k,g}^1 (e^{-\alpha_1 T_{1,k}} - 1) + (1 - \beta_{1,g}) (N_{1,k,g} + B_{1,k,g}),$$

where $N_{0,k,g}$ and $B_{1,k,g}$ are the initial contaminations of unit $k$ of genotype $g$ in colony forming units on the exterior and interior respectively. $\beta_{1,g}$ is the proportion of firmly-attached bacteria for genotype $g$ that we assume will remain on unit $k$ and $\alpha_1$ is the fraction of the free bacteria that transfer from the unit $k$ to the evisceration machine. The unit spends $T_{1,k}$ minutes in the evisceration machine. We assume $W_{1,g}^1 = 0$.

**Thermal inactivation process in hot water bath**

Assuming a line of units is submerged in a hot water bath after the first process, the solutions for the number of pathogen $X_g$ residing in the hot water bath or on the exterior of unit $k$ ($W_{2,k,g}^1$ and $N_{2,k+1,g}$) at the time of processing are:

$$W_{2,k+1,g}^1 = \left[ \beta_{2,g} N_{1,k,g} + W_{2,k,g}^1 \right]e^{-\gamma_{2,g}(1 - \alpha_{2,k}) T_{2,k}},$$

$$N_{2,k+1,g}^1 = W_{2,k+1,g}^1 \left[ \frac{\alpha_{2,k}}{\gamma_{2,g}(1 - \alpha_{2,k}) - \gamma_{2,k}} \right] \left( e^{-[\gamma_{2,g}(1 - \alpha_{2,k}) - \gamma_{2,k}] T_{2,k}} - 1 \right) + (1 - \beta_{2,g}) N_{2,k,g} e^{-\gamma_{2,k} T_{2,k}},$$

where $\alpha_{2,k}$ and $\beta_{2,k}$ have the same definitions as their counterparts for Stage 1 processing. The parameter $\gamma_{2,g}$ is the inactivation of genotype $g$ at temperature $C_k$. We assume $W_{2,g}^1 = 0$. The inactivation coefficient $\gamma_{2,g}$ is calculated by

$$\gamma_{2,g} = \frac{1}{D_{60,g}} \times 10^{\frac{60}{Z_{g}}},$$

where the D-value parameterisation is given for 60°C.

We assumed that as processing through the day continues the temperature of the hot water bath decreases over time from a baseline of 55°C as more material is mixed into the water. We also assumed a cyclical nature to the temperature over time, as operators monitor and
adjust temperature back upwards. Therefore, the temperature each unit $k$ is processed at is given by

$$C_k = 55 - mk + \left( 5 \times \sin\left(\frac{\pi k}{4}\right) \right),$$

where $m$ is the gradient coefficient of temperature decay.

---

**Figure 3:** Overall framework for hypothetical model. The processing steps are generalised from [23]. A unit $k$ enters process 1 (akin to evisceration) with various burdens of contamination from different genotypes (the model process is applied to each genotype independently, as shown by the faded out replications for Genotype 2 to and up). The number of each genotype on each unit is tracked as various cross-contamination and inactivation processes spread and reduce loads across the two processes (more detail on the notation used can be found in the appendix). A case of genotype $g$ may then occur according to a function of $D_{h,g}$, the dose of genotype $g$ ingested from unit $k$. 
**Extension to case reporting**

For simplicity we assume that there is no further cross-contamination, growth or inactivation along the food chain, and that a person consumes the whole of unit \( k \) in one meal. Therefore, the dose ingested by a person during one meal made with one of our processed units is the same as the number of bacteria within each population \( g \) at the end of stage 2 processing, that is, \( D_{k,g} = N_{2,k,g} \).

In the absence of any other data, we make the same simplifying assumption for the dose-response model as we did for the inclusion of separate genotypes above, that all sub-populations of \( X_g \) act independently. Therefore, the probability of infecting a single person with pathogen \( X_g \) is given by

\[
P_{k,g} \sim 1 - (1 - p_m g)^{D_{k,g}}.
\]

The likelihood of reporting an infection as a laboratory-confirmed case was assumed proportional to a virulence parameter \( v_g \). Therefore the probability of a reported case is

\[
L_{k,g} = P_{k,g} v_g.
\]

For each unit \( k \) on day \( d \) we determined whether a case was reported or not using \( R_{k,g} \sim \text{Binomial}(1, L_{k,g}) \). We summed the number of total laboratory reports over a week’s worth of production/consumption (assuming that the unit consumed is fresh produce consumed near the processing date), that is in one week the total number of cases would be given by \( \sum_d \sum_g \sum_k R_{k,g} \). This then replicated, in a simple fashion, the manner and incidence of reporting that we would expect for foodborne pathogens such as *Salmonella spp.* or *Campylobacter spp.*. We used this output as a way of replicating how public health authorities are often made first aware of a foodborne outbreak, where an increase in incidence is noted and investigated.

### 3.2.2 Genotype generation

The \( G \) genotypes will have varying phenotypes with regards to survival through the processing chain or in their ability to cause human infection and initiate severe disease. The generation of phenotype from genotype is a hugely complex area, and remains one of the key challenges for being able to use NGS and WGS data within risk assessment. We therefore created a simple dummy dataset to replicate five key genes that determine the phenotypes that map to the parameters in our model: two thermal tolerance genes...
representing the D-value at 60°C and the Z-value (TTD and TTZ); attachment efficiency (ATT); dose-response (PM) and virulence (V). In more detail:

ATT: Attachment efficiency gene – the proportion of strongly attached bacteria to the exterior of the product (and which will not be washed off).

TTD: D-value of genotype at 60°C on unit exterior and in water. Minutes taken to reduce bacterial subpopulation by 90%.

TTZ: Z-value of genotype g. The temperature required for one log10 reduction in the D-value. Used with the D-value to calculate the inactivation rate of genotype g at a given temperature.

PM: probability of human infection for one individual bacterium of genotype g. Used to represent the capability of human infection.

V: virulence factor representing severity of illness. We assume that the likelihood of reporting a case is directly proportional to the severity of the illness, hence $v_g$ can be used to assess the under-reporting factor for genotype g.

Each phenotype is subdivided into categories/alleles, for example the PM gene has alleles PM1-PM5. Hence if a case sample is whole-genome sequenced, we would know if that genotype possesses the PM1 or PM5 ‘allele’. The alleles of each genotype, and their associated phenotypic responses, are listed in Table 1, as well as the distributions from which these genotype parameter estimates are taken.

an example of the split of the phenotypic distribution into alleles is given for PM (see Figure 4). Each genotype was specified by its phenotypic characteristics, for example Genotype 1 may have ATT1, TTD3, TTZ4, PM1 and V5 phenotypes.
Figure 4: Example of how the range of the $pm_g$ distribution from the parameter estimation is broken up to categorise phenotype into genotype. So in this example, the PM1 “allele” has a probability of infection a person between $(0-2)*10^{-5}$, while PM5 has a range of approx. $(0.9-2)\times10^{-5}$. Hence the PM5 allele is more successful at causing human infection than the PM1 allele.
Table 1: Categorisation of phenotype into genotypes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
<th>Description</th>
<th>'Alleles'</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att</td>
<td>$β1_g, β2_g$</td>
<td>Attachment efficiency: the proportion of strongly attached bacteria to the exterior of the product (and which will not be washed off)</td>
<td>A1-A5</td>
<td>${0, 0.04, 0.08, 0.12, 0.16, 1}$</td>
</tr>
<tr>
<td>TTD</td>
<td>$D60_g$</td>
<td>D-value of genotype at 60C on unit exterior and in water. Minutes taken to reduce bacterial subpopulation by 90%.</td>
<td>TTD1-TTD5</td>
<td>${0, 0.4, 0.8, 1.2, 1.6, 2}$</td>
</tr>
<tr>
<td>TTZ</td>
<td>$Z_g$</td>
<td>Z-value of genotype g. The temperature required for a one log (factor of 10) reduction in the D-value. Used with the D-value to calculate the inactivation rate of genotype g at given temperature.</td>
<td>TTZ1-TTZ5</td>
<td>${5, 6, 7}$</td>
</tr>
<tr>
<td>PM</td>
<td>$pm_g$</td>
<td>probability of human infection for one individual bacterium of genotype g. Used to represent the capability of human infection.</td>
<td>PM1-PM5</td>
<td>${0, 0.3, 0.5, 0.7, 0.9, 2} \times 10^{-5}$</td>
</tr>
<tr>
<td>V</td>
<td>$v_g$</td>
<td>Virulence factor representing severity of illness. We assumed that the likelihood of reporting a case is directly proportional to the severity of the illness, hence $v_g$ can be used to assess the under-reporting factor for genotype g.</td>
<td>V1-V5</td>
<td>${0, 0.2, 0.4, 0.6, 0.8, 1}$</td>
</tr>
</tbody>
</table>
3.2.3 Output of the hypothetical food chain model

As discussed previously, the advantage of a hypothetical model is that we can know with complete coverage and precision the simulated processing conditions at any time point. This can also be related to the burden of the 100 genotypes on each food unit at all stages of the food chain. Hence the output of the food chain model is not simply how many infections were caused (which is not very interesting given the theoretical food chain construction), but rather the simulated data array that relates relevant physical and genotypic factors against whether an infection occurred. The schematic in Figure 5 illustrates the data which is stored within the model for each processed food unit, using the hot water bath as an example. We record the simulated temperature of the hot water bath for each food unit processed each day, coupled with detailed information on the initial and final contamination loads of each of the 100 genotypes on the food unit. Figure 5 displays an example of the output from the hypothetical food chain model (numbers do not necessarily reflect the final output used in the analysis).

Figure 5: Schematic representation of stored data array, using the scalding process as an example. The day and position of the unit processed form the first two dimensions of the stored data array, where the temperature of the scalding bath is stored at each position. The exploded diagram shows a number of bacteria of varying genotype \( g \in \{1..n\} \) on unit \( \{d,k\} \). We store the number of each genotype sub-population \( g \) on the unit at the end of scalding \( (N1_{d,k,g}) \), creating a 3-D array of contamination loads for each genotype on a unit \( k \) processed on day \( d \).
3.2.4 Parameter estimation

The parameter estimates for the hypothetical food chain model are shown in Table 3. Where possible, we generate the variation in the baseline parameter estimates from Swart et al.\textsuperscript{[23]} (we note where this is not the case). The parameter estimates from this paper are specific to \textit{Salmonella spp.} in pigs, but we use them simply as a way to generate some realistic values for processing parameters and a classical bacterial thermal tolerance or infection potential. We assume the variation in the parameter estimates represent the phenotypic variation across a species, derived from the specific genetic characteristics of each genotype.

3.2.5 Implementation

The model was developed in MATLAB R2015a, using the Statistics and Machine Learning toolbox (MATLAB R2015a, The MathWorks Inc., Natick, MA, USA). We ran the model for 1000 days, with 1000 units being processed each day. We included 100 baseline genotypes, representing the broad range of phenotypes as described by the parameter estimation.
Table 2: Example output from the hypothetical food chain model, which becomes the baseline dataset from which we draw upon to assess various surveillance systems. The final output of the model is whether or not a food unit (serving) resulted in a reported human infection (that is $R_{k,g} = \{0, 1\}$, where 1 is a reported case).

<table>
<thead>
<tr>
<th>Day</th>
<th>Unit k</th>
<th>$\alpha_{1k}$</th>
<th>$T_{1k}$</th>
<th>$\alpha_{2k}$</th>
<th>$D60_g$</th>
<th>$m$</th>
<th>Genotype g</th>
<th>$N_{0_{kg}}$</th>
<th>$B_{1_{kg}}$</th>
<th>$Z_g$</th>
<th>$pm_g$</th>
<th>$v_g$</th>
<th>$R_{k,g}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.0001</td>
<td>1.3</td>
<td>0.01</td>
<td>0.36</td>
<td>0.005</td>
<td>1</td>
<td>3.3x10^4</td>
<td>1.8x10^9</td>
<td>5.2</td>
<td>2.3x10^6</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>1.2x10^2</td>
<td>6.7</td>
<td>5.3x10^3</td>
<td>0.52</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.0002</td>
<td>1.4</td>
<td>0.05</td>
<td>0.23</td>
<td>0.007</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.2x10^2</td>
<td>2.1x10^9</td>
<td>5.8</td>
<td>4.3x10^3</td>
<td>0.45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Parameter estimates of the hypothetical food chain model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Distribution or equation</th>
<th>Unit</th>
<th>Value - mean (5th/95th percentile)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 g, \beta_2 g$</td>
<td>Proportion of firmly-attached bacteria</td>
<td>Beta(19,83)</td>
<td>-</td>
<td>0.16 (0.11; 0.22)</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$\alpha_1 k$</td>
<td>Proportion of bacteria transferring from food unit $k$ to machine</td>
<td>Uniform(0,0.0002)</td>
<td>-</td>
<td>1.0 $(0.1;1.9) \times 10^{-5}$</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$N_{0 k,g}$</td>
<td>Number of bacteria of genotype $g$ on unit $k$ at entrance to processing</td>
<td>Lognormal(3,3)</td>
<td>CFU</td>
<td>1,760 (0;2826)</td>
<td>Assumed</td>
</tr>
<tr>
<td>$T_1 k$</td>
<td>Time each unit $k$ spends in Stage 1 of processing</td>
<td>Uniform(1,2)</td>
<td>mins</td>
<td>1.5 (1.05;1.95)</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$B_1 g$</td>
<td>Number of bacteria of genotype $g$ extruded from within interior of unit $k$</td>
<td>Lognormal(3,3)</td>
<td>CFU</td>
<td>1,760 (0;2826)</td>
<td>Assumed</td>
</tr>
<tr>
<td>$\alpha_2 k$</td>
<td>Proportion of bacteria transferring from water to unit</td>
<td>Uniform(0,2) $\times 10^{-5}$</td>
<td>-</td>
<td>1(1.03;18.9) $\times 10^{-6}$</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$D_{60 g}$</td>
<td>D-value of genotype $g$ on unit at 60°C</td>
<td>Weibull(0.4,2)</td>
<td>mins</td>
<td>0.35 (0.09;0.70)</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$Z_g$</td>
<td>Z-value of genotype $g$</td>
<td>Uniform(5,7)</td>
<td>°C</td>
<td>6 (5.1;6.9)</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$m$</td>
<td>Temperature decay coefficient</td>
<td>Uniform(0,0.01)</td>
<td>1/k</td>
<td>0.005 (0.0005;0.0095)</td>
<td>Assumed</td>
</tr>
<tr>
<td>$T_2 k$</td>
<td>Time each unit $k$ spends in stage 1 of processing</td>
<td>Uniform(5,7)</td>
<td>mins</td>
<td>6 (5.10;6.90)</td>
<td>Fitted distribution to data from [23]</td>
</tr>
<tr>
<td>$p_{mg}$</td>
<td>Probability of infection from one organism of genotype $g$</td>
<td>Beta(4,999998)</td>
<td>-</td>
<td>4.00 $(1.37;7.60) \times 10^{-6}$</td>
<td>Assumed</td>
</tr>
<tr>
<td>$v_g$</td>
<td>Virulence/reporting factor</td>
<td>Beta(4,8)</td>
<td>-</td>
<td>0.33 (0.13;0.56)</td>
<td>Assumed to be around 1 in 3</td>
</tr>
</tbody>
</table>
4 Surveillance scenario analysis

4.1 Hypothesis and data-driven exploratory analyses

Due to the preliminary nature of the study, the analyses we conducted were driven by what data we thought might realistically be collected from a food safety surveillance system (both now and in the future) rather than derived from specific policy questions. The aim of these analyses is to show the potential for, and limitations of, combining genetic and epidemiological metadata, and to inform the direction of future research. Therefore, the surveillance scenarios chosen are necessarily exploratory and dealt with at a reasonably top level. However, we still wished to describe various plausible surveillance scenarios and show where on the spectrum of data collection (from WGS data only to complete knowledge of the food chain system) we could identify the "signal" of foodborne infection from the "noise" of the large stochastic variation in genetic and environmental factors from food unit to food unit.

4.2 Scenario 1: WGS data alone – early detection algorithm

We started by analysing the capabilities afforded by investigating WGS data alone. Given the paucity of data available from across the food chain, this is a similar situation to where we find current food safety surveillance systems. As a surveillance tool, the only practical use of this type of isolated surveillance data is to construct time series of infectious intestinal disease, using WGS to provide greater discriminatory power to identify outbreaks. Analysis of these time series over months and years allows the construction of early detection algorithms, such as originally described by Farrington et al. and others since [24-25]. These algorithms are used to detect anomalies in the time series of various pathogens, for example *Salmonella spp.* or *Campylobacter spp.*, which indicate whether more cases are occurring than usually expected at that time of year. Typically a moving average is adjusted for seasonality, and then confidence intervals are constructed. The upper confidence interval is used as an exceedance score: if reported cases exceed this score then a flag is raised to indicate the presence of an anomaly.

From the 1 million food units recorded in our simulated dataset, we constructed weekly totals of reported cases for each genotype g. As there is no seasonality in our
hypothetical model (negating the need to calculate a moving average), we use the 95th percentile value of the distribution of overall weekly reported cases across all genotypes as a very basic exceedance score. This also avoids the need to choose a particular type of early detection algorithm to apply. We subsequently modified the baseline model to generate two scenarios: one where a random genotype gains 100x its previous efficiency to cause human infection, and another to investigate the effects of a systematic error in processing by introducing an average 2°C drop in Stage 2 water temperature over a two-week period. We thus interrogated the time series to observe how often a flag would be raised and whether such a WGS early detection model could detect these changes from the baseline.

4.3 Scenario 2: WGS data alone - WGS collection during microbiological sampling at processing

One of the major problems with food chain surveillance is that WGS data are only consistently collected at the human end of the food chain, with little or no characterisation of whole genome sequences across the food chain. It is therefore almost impossible to identify, for example, survival genes if there is no concept of what strains did not survive the processing chain. We therefore investigated the collection of WGS data in the food chain itself.

Microbiological sampling is a legislative cornerstone of food safety within the European Union. Samples taken during processing of food (e.g. samples from pig or chicken carcasses tested for *Campylobacter spp.* or *Salmonella spp.*) allow a Food Business Operator and enforcement agencies to monitor the performance of the system to produce safe food. This microbiological sampling is legally required and hence presents a consistent and robust set of data with which to compare microbiological status of food during processing with human infection. Currently, microbiological criteria are set at species level, and so there is no requirement to assess the genomic structure of any pathogen present by NGS or WGS. However, if WGS technologies continue to become cheaper and are used on microbiological sampling isolates, we could potentially have a bounty of well-structured WGS data at the processing stage with minimal additional infrastructure and cost. Therefore we tested the extra capability that WGS could bring by being able to link genotype at processing and human infection stages. We applied a five- samples-per-week scheme (comparable to the microbiological sampling scheme for Salmonella in pigs), and assumed each microbiological sample would pick up any genotype on a sample with more than 100 Colony Forming Units (CFU) on a food unit.
Therefore, over a period of 1000 days (about 143 weeks), there would be 5000 individual samples tested (we randomly pick 5 samples from across every week). We simply plotted the rate of reporting across genotypes against the burden of contamination by genotype.

### 4.4 Scenario 3: Classical statistical analysis of epidemiological metadata and genomic data

One of the most problematic issues in food safety surveillance is being able to backtrace an outbreak along the food chain to identify the original source of the outbreak. In the third scenario we therefore introduced the theoretical capability to track and trace along the food chain (e.g. by RFID tagging), allowing a reported case to be allied to a sample at the stage of microbiological sampling. This would represent a significant increase in fidelity and resolution of epidemiological data, and would potentially allow us to start to look at interdependencies between epidemiological and genomic factors.

We assumed a perfect implementation of track and trace: that is, for each reported case we can track the food unit consumed back through the food chain and identify the processing conditions (in this case the temperature of the hot water bath) at the time of processing that specific food unit.

We first conducted a traditional statistical analysis of such a dataset including positive/negative record of a case, along with associated explanatory variables such as processing temperature and the genotypic identity of the infecting organism. Whether a unit leads to a case or not is a categorical variable: we therefore used logistic regression to estimate the probability of a case given the temperature and phenotypic identity of the organisms contaminating that unit. We specified a binomial model with a logit link function. That is, we determine the relationship between the binary response of a case (or not) to the unit processed, the temperature of the hot water bath during processing and the genetic profile of the genotype \( g \) (e.g. genotype 1 may have the profile \( G1 = \{ATT1, TTD3, TTD1, PM1, V3\} \)).

Based on preliminary results (see Section 5.2.3), we also investigated how the logistic regression model would perform if the assumed WGS microbiological sampling system was improved by i) assuming each sample is enumerated, so that we also have a measure of contamination burden as well as presence, and ii) taking 250 (rather than five) samples are taken per week, to produce more cases that can be traced back to
food units that were sampled during processing. Clearly, this surveillance system would be superior to anything currently available anywhere in the world, and would certainly be very expensive. However, the implementation is of interest if only for demonstration purposes, as knowing what knowledge gains such a system could bring can help us to establish where to prioritise data collection activities. For example, could simple enumeration of samples provide a bigger jump in knowledge than genomic information?

4.5 “Big” epidemiological data - track and trace individual food units

Finally, we used the increasingly popular approach of machine learning to investigate whether, given sufficient data, a predictive human infection model could be developed, using the same surveillance scenario as described for Scenario 3. We do not contend that such a predictive model should (yet) be used as a decision-making tool, but we wanted to explore the potential of predictive analytics methods to strip away the noise of the food system and begin to understand the complex interdependencies between genetics and physical systems.

Such a surveillance system would therefore give us a dataset that identifies the processing characteristics of food units that caused reported infections, with an abattoir dataset that will tell us the genotypes on the food unit as it was processed. So for a subset of this dataset, we know which food unit caused a reported infection, the processing conditions and the genotype burdens contaminating it.

There are good, accessible reviews of machine learning in the context of genomics or epidemiology [26–28]. The basic principle of machine learning is that there are enough data to split a dataset into training and evaluation datasets. A predictive model is fitted to the training data, and then evaluated against the evaluation dataset (using some statistical measure of fit, for example Mean Squared Error, MSE, which we used in this study).

A very popular and relevant branch of machine learning is the use of decision trees, where the predictor variables are partitioned using a series of rules to identify regions that have the most homogeneous response to predictors. Decision trees are advantageous from a number of aspects important for food safety because they can i) cope with a variety of data types, ii) the method requires no pre-processing of data and iii) they can deal with missing data points [26]. These are all important considerations for food chains where currently the data collected are of varying provenance and quality.
The hierarchal structure of decision trees (where one input variable depends on inputs higher up in the tree) means that interactions between variables are automatically captured.

We used MATLAB's machine learning toolbox to build a predictive algorithm. A major decision when growing a predictive decision tree is to set the “leafiness” or depth of the decision tree. Deep trees with many branches may very accurately reflect the training data, but are potentially over-fitted, so that the tree algorithm is not successful at prediction when applied to other (evaluation) datasets. Indeed, our preliminary analyses showed this to be the case (not shown). We therefore applied a boosted decision tree, a popular type of Classification and Regression Tree (CART) algorithm [29] that is typically a more accurate method of producing a predictive algorithm. Boosting involves fitting a series of ‘weak learners’ (simple decision trees) to the training data. Briefly, each datapoint is first weighted equally and the first simple decision tree is fitted. The only requirement for this (and every) tree is that the prediction is better than random chance. The weights of those datapoints incorrectly classified by the first classification tree are increased so that the next decision tree fitted is focused more on fitting the incorrectly classified datapoints. This continues for as many trees as specified. A score is assigned to each classifier, and the final classifier is defined as the linear combination of the classifiers from each stage [30]. Other critical decisions when developing a boosted decision tree is the number of splits to include within each individually fitted tree (one split being equivalent to two branches within a tree). Another parameter to set is the “shrinkage” learning rate, $\lambda$, where $\lambda \in (0, 1)$. The smaller the learning rate, the less weight given to each sequentially-fitted tree; this increases the time the model takes to converge on a solution, but generally increases the accuracy of the eventual solution [29,30].

In our case, we have what is a relatively small, simple dataset (one response variable and six main predictor variables), but one which is characterised by a common type of problem termed class imbalance, where we have an over-abundance of non-cases and a small minority of units that lead to cases. We chose the RUSboost method [31], as this is designed to handle the problem of class imbalance (through Random Under-Sampling - RUS) as well as increasing the predictive performance of the algorithm through boosting.

Finally, we evaluated the dataset using cross-validation. Cross-validation involves partitioning a sample of the data into complementary subsets, performing the analysis on one subset (the training dataset) and validating the analysis on the other subset. To
reduce variability multiple rounds are performed using different partitions (we used five 'k-folds', essentially splitting the data into five roughly equal partitions). The validation results (i.e. the MSE values) are averaged over the rounds.

The following outlines the chronology of the machine learning method we implemented:

- Construction of the simulation dataset. We used the subsample of food units that were sampled under microbiological sampling. Based on the initial analysis of the raw dataset described in Section 5, we chose to include the following variables in the machine learning dataset: i) the contamination of the food unit at the end of Stage 1 processing, ii) the temperature of the hot water bath, and iii) the D60, Z, PM and V alleles as our predictor variables. The binary classifier of whether a case was reported or not was the response variable.
- Imported this dataset into MATLAB’s machine learning toolbox.
- We chose to implement the RUSboost decision tree method, for the reasons outlined above.
- Set a range of values for the number of splits and learning rate. We investigated split number of 1, 4 and 7, and a learning rate of 0.1, 0.5 and 1. We therefore generated nine predictive algorithms.
- We implemented five rounds (k-folds) of cross-evaluation, and used the MSE as the performance metric (the lower the MSE, the better the predictive algorithm).
- We analysed the predictive capacity of the 'optimal' predictive algorithm (as assessed by comparing MSEs in the step above).

5 Results

5.1 Exploration of simulated data array

The data array produced from the hypothetical food chain simulation gives, for each simulated food unit, the physical and genotype parameters and whether an infection with pathogen \( g \) occurred given the consumption of that food unit. This data array provides the foundation for the surveillance system analyses that follow. We first analysed the raw data array to better understand the relationships between parameters and reported infections. We plotted daily parameter averages against average number of reported cases per genotype for phenotype indicators (Figure 6) and average cases per day for processing parameters (Figure 7). We placed contamination rates in the processing parameter figure as these parameters are generated independently of the genotype. These simple analyses highlight that of the phenotypic indicators, only the thermal inactivation parameters whilst bacteria reside on the unit (\( D_{60g} \) and \( Z_g \)) and the infection and reporting parameters (\( pm_g \) and \( v_g \)), are firmly correlated with reporting rates. For the processing parameters, the temperature of the food unit appears to be a very strong driver of infection. The response to initial and post-
process 1 contamination rates \((N0_{k,g} \text{ and } N1_{k,g})\) was similar and showed a modest positive correlation (only \(N1_{k,g}\) shown in the figure). While these simple linear fits are not robust statistical analyses (for example, in reality we would expect to see at least some non-linear relationships, especially for those parameters driving cross-contamination), the visualisations do provide some good clues as to what parameters should drive infection and reporting rates. Therefore, we imported into the machine learning toolbox the four genotypic parameters with reasonable correlations to reported infection \((pm_g, v_g, D60_g \text{ and } Z_g)\), plus the contamination at the end of Stage 1 processing and the temperature of the processing conditions during Stage 2 processing.
Figure 6: Scatterplots showing the relationships between phenotypic indicators and the number of reported cases per genotype over 1000 days. Blue lines represent least squares regression fits with 95% confidence intervals (red dash-dot lines). For clarity we show a common scale across the panels: some confidence intervals are therefore not shown (for example the lower confidence intervals that fall below zero). We also separate the effect of D60g and Zg in Stage 2 water and on the food unit: this demonstrates the effect of physical context: inactivation in water is relatively unimportant, but inactivation on the food unit is certainly correlated with infection rate.
Figure 7: Scatterplots showing the relationships between the processing parameters (averages over day) and the average number of reported cases per day. Blue lines represent least squares regression fits with 95% confidence intervals (red dash-dot lines). For clarity we show a common scale across the panels: some confidence intervals are therefore not shown (for example the lower confidence intervals that fall below zero).
5.2 Scenario analysis

5.2.1 Scenario 1: WGS data alone – early detection algorithm

The weekly totals generated from the simulated data array are displayed in Figure 8. We arbitrarily set the exceedance score to the 95th percentile of total weekly cases (41 per week), such that there are roughly 2-3 weeks per year that exceed the exceedance score. In our baseline dataset, these flags are false positives, as there is no change from the underlying trend.

To investigate the sensitivity of this basic early detection algorithm we modified the baseline model to mimic two types of outbreak: i) a genotype gaining a gene or plasmid that makes it far more efficient at human infection, increasing the probability of any one organism causing infection ($p_{m_g}$) by either 10x or 100x, and ii) a processing failure that leads to a drop in temperature of the critical control point second processing stage. We then observed how our simple early warning algorithm performs in detecting these types of changes.

One genotype gaining greater 10x human infection efficiency was unlikely to register within overall weekly reporting of cases; however an increase in 100x may be noticeable (see Figure 9). For example, a genotype with 100x efficiency would make that genotype around 10-100x the efficiency of the previous maximal efficiency genotype. More detailed analysis by the responsible authorities may well detect an anomaly, even if an outbreak alarm is not triggered due to the relatively low increase in overall cases. However, it is likely that this anomaly may be just one of many, potentially swamping any outbreak signal with noise.

A similar pattern emerged when we temporarily dropped the average temperature of the second processing step (arbitrarily taking weeks 73-80). An average 1°C drop in temperature led to a clear increase in cases (Figure 10), but may well be mistaken for a false flagged event if further information (e.g. follow-up questionnaires for reported cases) were not sensitive enough to identify the processing establishment as the common link. A 2°C drop led to a bigger, more sustained exceedance of the threshold, and would therefore increase the likelihood of detecting the processing failure. Both temperature failure scenarios could therefore be detected by an early warning system, but it may require some significant detective work in the 1°C scenario. An average of 1-2°C is a large drop in scalding temperature, and it is likely such drops in temperature would start to affect the quality of the food product. Hence, the processing establishment may well detect and rectify the problem themselves much sooner than the surveillance system could.
Figure 8: Weekly cases as generated by the hypothetical food chain model. Weekly cases by the 100 unique genotypes are stacked to produce overall numbers. The grey dotted line represents the 95th percentile, used as an arbitrary exceedance score.
Figure 9: Change in overall weekly reporting if one random genotype gains efficiency to cause human infection, increasing the probability of infection from one organism by 10x-100x. Because the signal from the one genotype is swamped amongst many other genotypes, it needs a huge increase in infection efficiency (100x) to consistently breach the exceedance score.

Figure 10: Weekly overall reported case totals: grey - baseline; red - average 1°C drop in processing temperature for eight weeks (Weeks 73-80); blue - average 2°C drop in processing temperature, same time period. Grey dotted line represents exceedance level: in the absence of more information, the 1°C drop may be missed or put down as a false positive; the 2°C would almost certainly be identified as an outbreak.
5.2.2 Scenario 2: WGS data alone – addition to microbiological sampling criteria

Assuming that 250 microbiological sampling criteria samples are taken every week, we plotted the number of genotype \( g \) isolated during microbiological sampling over the 143 weeks against the number of human cases attributed to genotype \( g \) (Figure 11). The scatterplot clearly shows the weak relationship between WGS information at the abattoir and reported cases; the wide regression confidence interval supports this conclusion. We are clearly missing some crucial information.

![Figure 11: Scatter plot of the total number of genotype \( g \) samples isolated during process microbiological sampling over 143 weeks of processing (5 samples every week). The blue line is the least squares regression fit, the red dash-dot line the upper 95% confidence interval. We have removed the lower confidence interval for clarity.](image-url)
5.2.3 Scenario 3: Classical statistical analysis of both epidemiological and genomic data

The logistic regression output is shown in Figure 12. Overall the logistic regression model did reasonably well, i) identifying the PM gene as being strongly correlated with whether a case is reported, and ii) that there is an inverse relationship between temperature of unit processing and whether that leads to a human case. However, the logistic regression model predicted a strong inverse relationship between the virulence/reporting gene $v_g$, which is inconsistent with how the model works – the reporting factor is directly proportional to the number of cases. The lack of accuracy is probably due to the large number of samples that would be required to generate a sufficiently large number of human cases for the method to be robust – many of the parameters are linearly dependent (that is, one category of a predictor variable is always associated with no cases). This linear dependence problem is solved by raising the number of samples taken per week to a theoretical number of 250. This then generates enough human cases from sampled units to remove the linear dependence issue. The resulting logistic model from this higher-sampling programme produces a much more reliable model (see Table 4).

The main reason for the linear dependence in this type of assessment is that, overall, in this model and in reality, the rate of conversion from contaminated unit to infection is very low. This presents a problem for interpreting any data from microbiological sampling, as even with the 5000 samples generated over roughly three years of operation through our hypothetical model, the vast majority (if not all) of these samples will not result in a human case. We therefore need to be able to relate a large number of cases to data along the chain if we are to have a real chance of maximising the analytical capabilities of a surveillance system.

For a genuine assessment of risk we must also know something about the concentration of the organism on a food product (as ultimately all infection response is exposure and dose-dependent). We therefore included the concentration of the subtype g at sampling into the logistic regression model (see Table 4), which becomes the largest parameter estimate of the model. This simply emphasises the importance of considering the exposure component
of the risk equation as well as considering the infection potential of the organism.

Figure 12: Results for logistic regression of link between microbiological sampling criteria samples and human cases (using five microbiological samples taken from the processing unit each week). Due to the low rate of infection from contaminated units, not many sampled units result in cases. Therefore, the results of the logistic regression model are not quite reliable; most parameters are well-estimated, but the reporting factor is clearly wrong, with higher virulence genotypes (e.g. V4) having a lower probability of a reported case than low-virulence genotypes such as V1 and V2.
Table 4: Logistic regression coefficients in various scenarios: i) using standard 5 samples per week criteria for microbiological sampling, ii) using 250 samples per week, and iii) adding in knowledge about the concentration of each genotype g on unit k at sampling.

<table>
<thead>
<tr>
<th>Logistic regression model coefficients</th>
<th>Five samples per week</th>
<th>250 samples per week</th>
<th>250 samples per week + concentration parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant, $\beta_0$</td>
<td>5.51</td>
<td>2.53</td>
<td>-4.61</td>
</tr>
<tr>
<td>$B_{\eta_1}$ (concentration)</td>
<td>-</td>
<td>-</td>
<td>2.23</td>
</tr>
<tr>
<td>$\beta_{\tau_1}$ (temperature)</td>
<td>-0.25</td>
<td>-0.22</td>
<td>-0.25</td>
</tr>
<tr>
<td>$\beta_{\text{eff}}$ (attachment)</td>
<td>-0.02</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>$\beta_{D_{60}}$ (D-value)</td>
<td>0.85</td>
<td>-0.14</td>
<td>-0.06</td>
</tr>
<tr>
<td>$\beta_{Z}$ (Z-value)</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>$\beta_{\text{pm}}$ (Infection)</td>
<td>0.75</td>
<td>0.37</td>
<td>0.44</td>
</tr>
<tr>
<td>$\beta_{\nu}$ (Reporting)</td>
<td>-1.12</td>
<td>0.38</td>
<td>0.43</td>
</tr>
</tbody>
</table>

5.2.4 Scenario 4: “Big” epidemiological data - track and trace individual food units

Using the Mean Squared Error (MSE) as described in Step 5 of the machine learning method, the optimal predictive model was one with a maximum of seven splits per tree, and a learning rate, $\lambda$, of 0.5 (see Figure 13). There is a slight increase in performance between three and seven splits, but little was added to the performance by increasing splits above seven (not shown). We chose to use 300 decision trees, but there was little improvement in the MSE above 200 trees.

Most trees are dominated by nodes splitting the level of contamination at the end of Stage 1 processing ($N_{1}k,g$), followed by the temperature of the Stage 2 hot water bath ($C_{\nu}$) (not shown). Of the genotypic factors, the PM allele is the most common, but appears relatively infrequently compared to the physical parameters. This strongly suggests that contamination burden (regardless of genotype) is a key influence of whether or not infection is reported. Indeed, the impact of contamination level and temperature is clear in the decision surface produced by using the predictive model to predict classification of reported infections for various contamination levels and temperatures (top left panel of Figure 14).
Figure 13: Comparison of the performance of the RUSBoost algorithm in predicting the reporting of human cases for varying number of splits in the trees (MaxNumSplits), the number of fitted trees in each algorithm (x-axis) and for learning rates 0.01 - 1. We used the cross-validated Mean Squared Error (MSE) as the measure of performance. Using seven splits produced the best-performing algorithm; there was little performance increase using more than 200 trees per algorithm.

Figure 14: Decision surfaces for selected comparisons. Those combinations of parameters classified as no reported infection are coloured blue, those classified as reported infection are red. Clockwise from top left: log contamination of exterior at the end of Stage 1 processing (N1) vs temperature of hot water bath (C); N1 vs. the allele associated with probability of infection, PM; C vs. PM; and finally PM vs. the virulence allele, V. When not included in the comparison, the following parameters were set at the median values: PM = 2, V = 2, D60 = 1, Z = 1. We set N1 = 1.4x10^4 CFU and C = 52.8°C to produce boundary conditions that allowed us to show the impact of the PM and V alleles in the top-left and both bottom panels.
As we would expect, combinations of higher contamination and lower temperatures drive the decision surface classified as reported infection. The top-right and bottom-left panels of Figure 14 illustrate the importance of 'physical' parameters (contamination level and temperature) relative to the most influential genotypic factor, the probability of infection. The PM allele makes little difference to the classification of a unit, except when right on the cusp of the decision boundary between contamination level and temperature. For example, only fitter PM and V alleles are classified as reported infections when on the decision boundary of contamination level and temperature (using the decision boundary that occurs at $1.4 \times 10^4$ CFU and 52.8°C) (see bottom-right panel). Thus we could hypothesise that there is a large decision space where only classic parameters such as contamination level and critical control points such as scalding water temperature are important. Genotypic alleles are therefore only important on the cusp of decision boundaries. Given food hygiene processes such as Hazard Analysis and Critical Control Points (HACCP), these decision surface boundaries are hopefully most commonly approached when there is a failure in the food production chain.

One of the key aspects of machine learning is the division of the original dataset into training and evaluation datasets. The training dataset is used to develop the boosted decision tree solution, and the evaluation dataset is used to test the predictive power of the predictive algorithm. We therefore assessed the performance of the optimal predictive algorithm to predict whether a case is reported using the evaluation dataset set aside from the original dataset. This information is displayed using a confusion matrix (Figure 15), which describes the number of false positives, false negatives etc. From these figures we can work out the sensitivity of the predictive model, that is $Se = \frac{1102}{(1102 + 315)} = 0.78$, and the specificity, $Sp = \frac{1,571,036}{(1,571,036 + 82,378)} = 0.95$. Therefore, our predictive model would be relatively good at detecting combinations of environmental and genomic characteristics that lead to cases. Hence, if such big data were available, the authorities could very well use such techniques to identify and monitor potentially high-risk scenarios. However, our results exhibit a classic example of the problem of looking for a rare event: we can detect roughly three out of four cases, but the total predicted cases are dominated by false positives due to an over-abundance of negative results. Therefore the predictive power of the model is poor: the positive predictive value (PPV) is only 0.01.
6 Discussion

6.1 Discussion of results

Our hypothesis stated that to generate genuine knowledge from high fidelity genomic data for food safety surveillance we should also link it to relevant, high-fidelity epidemiological metadata, using data from across the food chain. Our analyses therefore focused on identifying where a useful step-change in knowledge may be generated from ever more detailed genomic or epidemiological metadata. In short, our preliminary conclusions are that tracking and tracing food units at as high a resolution as possible is necessary to genuinely transform (genetic) data into knowledge of epidemiological risk. However, the majority of extra knowledge gained from tracking and tracing was due to being able to determine the success of the food chain in controlling contamination burden, regardless of the genetic data associated with a case or food unit. Only on the cusp of decision boundaries (that is where small changes in either contamination or the efficiency of the food chain to mitigate that contamination tip the balance in favour of infection or no infection) did we see the influence of genetic factors.

We first tested a surveillance scenario that looks much like what is present in the UK now: using genotyping information from reported cases to interrogate weekly cases trends. We showed that changes in the baseline system have to be quite large to be detected - a
consequence of the huge variation that exists in a food production system (even from our simple food chain model). This means that the exceedance score has to be set reasonably high to avoid a burdensome false positive rate, otherwise authorities would expend precious resources on following up what are in all likelihood spikes in the natural trend rather than genuine outbreaks. Without further contextual information the resolution of the prediction being made of these early warning systems is relatively low. It simply states “there may be an outbreak”, but potentially significant epidemiological investigations may be needed to establish if there is an outbreak, and if so what is the origin. These algorithms (of themselves) do not answer the questions “why did it happen?” or “what will happen?“.

Currently, WGS has shown its value in some outbreaks where the same genotype can be linked to a single source and may well accelerate source identification or provide robust evidence for one particular contamination source over others [12-14]. However, the main purpose of outbreak investigation is to control and mitigate the source of infection; while some useful insights may be gained by these investigations, they do not provide a standardised and robust way to link genomic and epidemiological data, and indeed the epidemiological data collected can be inconsistent and incomplete due to the nature of backtracing along a complex and dynamic system.

Our second surveillance system included the data from the first scenario (reported case WGS data), plus WGS information from the processing stage. This would be an advance over most current surveillance systems across Europe, but is not that unrealistic, as we assumed we could access samples from mandated microbiological sampling. However, there was no clear correlation between rates of WGS at processing and reported cases (see Figure 11). This is not unexpected, given that such a system would be based on presence of a genotype, rather than burden. Contamination levels (in this model, and observed in many field studies of food and animal samples) regularly vary across many orders of magnitude: if contamination level is not enumerated, this wide variance is very likely to overwhelm any genetic signal with noise.

Of wider interest is that implementing WGS in microbiological sampling schemes produces data that could be used within the common and popular source attribution models based on the seminal work by Hald et al [32]. Their model is based on the principle of being able to construct joint distributions for a couple of key parameters that describe the food-dependent and bacteria-dependent factors. However, one of the major issues with this type of Bayesian model is the sensitivity of the final joint probability distribution to the choice of prior information, and commonly the lack of enumeration data as mentioned above [33]. This is critical as we are currently very poor at being able to generate prior estimates for these
parameters (especially contamination burdens) because food surveillance systems do not collect data in sufficient detail from across the food chain. While the Hald models provide valuable information for source attribution and the prioritisation of resources to the most problematic sources, our hypothetical analysis shows that better contextual data from across the food chain, especially on contamination burden, is still necessary, especially if these methods are to be used with WGS information (either to better populate the Hald models, or to use new approaches). Similar conclusions were reached by the authors of a Danish study investigating the use of MVLA methods for Salmonella attribution [34].

The most important aspect that increased the fidelity of metadata was to add the ability to track and trace individual units. Without this detailed track and trace data, the noise of even our simple food chain overwhelmed any patterns that we may have been able to identify. Logistic regression was not possible until the number of microbiological samples from processing was increased from five to 250 (otherwise it was likely no reported cases were generated from the sampled units). The 250-sample logistic regression model was reasonably accurate and reflected the dynamics we expected to see, although the addition of contamination load highlighted the importance of this parameter in driving risk.

The machine learning example (Scenario 4) used the same data as the logistic regression model. Given the use of decision trees, we were also able to identify numerous interactions that lead to cases, although the majority of decision trees simply expressed the relationship between contamination level at the end of Stage 1 processing and the temperature of Stage 2 processing. The contribution of genetic information was limited in comparison, although the RUSBoost algorithm did identify that more efficient infection potential (PM3-PM5 alleles) and higher virulence (V3-V5) were associated with cases on the decision surface boundary between contamination level and temperature (see Figure 14). Thus, given appropriate surveillance data, machine learning does provide a potentially attractive option for analysing food chain data, with the ability to determine the importance of interactions between novel genes/alleles and the processing environment. A further research project could concentrate on an area where we have real, relatively good cross-food chain data, for example a big integrated supply chain.

The predictive model was reasonably sensitive if all information was known about a food unit, but the positive predictive value of the model was poor due to the rare nature of foodborne infection. In our hypothetical model, there were around a thousand units consumed for every reported case. Thus, even though the predictive model is quite specific (95% specificity), the number of false positives dwarfs the number of true positives. This is the well-known issue of class imbalance and occurs in any surveillance system dealing with
relatively few numbers of positives. Hence, we do not suggest that such a predictive model is currently used to “predict” outbreaks or the burden of foodborne illnesses, but rather is used to identify scenarios of combined environmental and genomic characteristics that may prescribe a higher risk of infection in the consumer population. This could supplement current early detection algorithms, either to improve the predictive power of such algorithms, or to provide extra context to qualitative horizon scanning efforts, for example to assist in the identification of the provenance of an outbreak.

6.2 How “big” epidemiological metadata could be used to supplement current food safety surveillance

The advantage of big data is that, given relevant data sources, we should be able to collect enough data to fully represent the true distribution of variables and outputs, something that we cannot be confident about current data availability. This in itself would be a big step change in the field of food safety, where analysts commonly have to work with poor, incomplete data with small sample sizes. Paying significant attention to data generation, curation and analysis through the application of logistical and technological advances to both genomic and supply chain data would undoubtedly improve the representativeness of collected data compared to the current situation. This improvement in data curation and collection is a first priority above and beyond attempting to build models that can predict outbreaks or foodborne illness.

Based on these initial analyses, we can suggest how “big” metadata might supplement and improve upon the current framework for food safety surveillance (Figure 16). Briefly, the addition of contextual metadata would allow us to better inform food safety policies at the national level, proactively identifying higher-risk scenarios that should be controlled. Monitoring of trends in near real-time would also allow extra contextual information to be provided in the case of an alert of a potential outbreak and better direct epidemiological investigations.
Figure 16: Top-level framework for the inclusion of high-fidelity metadata into a food safety surveillance system. The black-text boxes represent current activities, the grey-text boxes where additional metadata may be included to identify higher-risk scenarios and hence inform both the longer-term policies to reduce overall foodborne illness, and the reactionary outbreak response, by providing extra contextual information to the current early detection algorithms. Dotted lines represent indirect contributions.

6.3 Implementation (barriers and practicalities)

There are of course large barriers to the implementation of a system that can track and trace food units or batches. A survey of poultry and pig producers conducted during the study showed that the ability to track and trace food products is varied. Some companies do have centrally-collated electronic databases of sampling results and processing data, but some only have locally-stored paper records, which may or may not be interrogated further. There would also likely need to be investment in the technologies required to transmit, store and collate the large volumes of data generated by real-time sensors. Hence, the introduction of RFID tagging would have to be shown to be cost-effective. Even if RFID tracking and tracing
was introduced, under current arrangements this information would not be immediately available to local or national authorities. Hence, there are several issues around data generation and access that would have to be surmounted first. One attractive area of research would be to establish exactly what fidelity of track and trace data may both improve food safety surveillance but also production performance? If these RFID technologies are shown to be beneficial in improving production efficiency, then food chain companies may well take up the technologies for their own purposes, with food safety surveillance being improved as a byproduct.

Big data approaches are gaining in popularity, and are hailed as a way to improve decision-making and increase competitiveness in the market [35]. They have been applied in a variety of related fields in genetics (genomics, proteomics), but detailed metadata are rarely taken into account, and as such methods tend to focus exclusively on the information that can be gained from the genetic data alone (presumably because there is usually a lack of complementary metadata). However, as we have shown through both classical statistical and big data approaches, we should not ignore epidemiological and contextual data. The burden of contamination and the processing environment, rather than the genetics of a novel strain, are still the most fundamental aspects for control of foodborne pathogens. A novel, virulent foodborne pathogen cannot cause illness if exposure does not occur, or doses ingested are too small to cause infection. As always, we must consider both the opportunity for exposure and the capability of infection/virulence of disease.

The research presented here is preliminary and theoretical, but directed at the question of how big metadata would support the development of more sensitive surveillance systems based on WGS technologies. The surveillance scenarios, and their analyses, are necessarily exploratory and presented at a top level. Detailed study of the application of big data to specific food surveillance policy questions will come later, but hopefully this research will inform the direction of that research. A lot of research and development is required between now and the application of the machine learning methods to directly influence operational activities and epidemiological investigations. This includes: ongoing research into translating genomic data into risk-based knowledge of genotype through matching the most relevant, high fidelity epidemiological data to genomic data; researching other machine learning methods to develop and improve the predictive/insight capability of such methods; liaising with industry to determine the reception of companies to employ (and share the results of) big data systems, and to conduct operational research activities to identify the optimum real-time data gathering processes that allow industry to gain better productivity and that allows both industry and government to identify the most cost-effective measures for food safety.
6.4 Next steps and recommendations for further research

The research presented here is theoretical but directed to the question of how big metadata would support the development of more sensitive surveillance systems based on WGS technologies. There is much research and development in between now and the application of the machine learning methods to directly influence operational activities and epidemiological investigations. There are a number of steps to achieve the type of system proposed in Figure 16. They include:

1. Identifying the quality, quantity and accessibility of current data within food chains;
2. Further analytical research to investigate the most appropriate big data methods, and the most relevant metadata to use with them;
3. Ongoing research into translating genomic data into risk-based knowledge of genotype;
4. Consultation with industry to determine the reception of companies to employ (and share the results of) big data systems;
5. Operational research activity to identify i) the cost-effectiveness of a WGS and big data enabled system, and ii) determine the most data and cost-efficient infrastructure to collate the required data;
6. Pilot trials to introduce such systems (perhaps at a regional level or with specific producers).

6.5 Recommendations

Immediate

6. Meet with Warwick PL to discuss potential synergies and attend their workshop to present results;
7. Through collaboration with Warwick, produce a combined set of benefits and recommendations from both projects;
8. Refining our working hypothesis to better test it – for example can we specify in greater detail what level of track and trace fidelity is required?
9. Further investigation of proof-of-concept machine learning techniques, for example:
   a. Conducting machine learning analysis on a more complicated hypothetical model;
b. Other predictive analytic methods, such as neural network analysis, which could be used to identify risky combinations of genotypes and production factors.

10. Identify 3-4 areas of research where combination of real WGS data (perhaps from GMI) and big data would be beneficial.

Medium-term

5. Development of Performance Indicators for food safety surveillance (FSA strategy objective);
6. Further investigation of relevant metadata (for GMI) to enable measurement of Performance Indicators;
7. Develop infrastructure required to measure and assess Performance Indicators;
8. Discussion with industry over use of WGS and "big data" technologies, including track and trace, generation of networked metadata, and any access issues.

Long-term

5. Implementation of new logistical technologies into food production chain;
6. Agreement over right level of discrimination for use of WGS in surveillance systems and associated research;
7. Central data collection of metadata into GMI or a compatible database;
8. Implementation of new surveillance algorithms using both case data and epidemiological data.
7 References


## Glossary and abbreviations

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backtracing</td>
<td>Typically carried out during an outbreak investigation, where food leading to cases is backtraced along the food chain to establish the origin of an outbreak.</td>
</tr>
<tr>
<td>Boosting</td>
<td>A machine learning technique for boosting the predictive power of algorithms by combining many smaller models. Typically a more accurate way of producing a predictive algorithm.</td>
</tr>
<tr>
<td>CCP</td>
<td>Critical Control Point</td>
</tr>
<tr>
<td>CFU</td>
<td>Colony Forming Unit</td>
</tr>
<tr>
<td>Decision tree</td>
<td>A visual method of classifying outcomes by inputs.</td>
</tr>
<tr>
<td>Early detection algorithm</td>
<td>Common time-series analysis method used to detect anomalies in reporting trends. Used to guide outbreak investigations.</td>
</tr>
<tr>
<td>EFSA</td>
<td>European Food Safety Authority</td>
</tr>
<tr>
<td>(Epidemiological) metadata</td>
<td>Any other data that may accompany genetic data. Epidemiological metadata can simply be the species or type of sample from which a micro-organism was isolated, through to detailed information on the history of the product consumed (e.g. temperature at which the product was processed).</td>
</tr>
<tr>
<td>HACCP</td>
<td>Hazard Analysis Critical Control Point</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>Machine learning</td>
<td>A method for producing predictive models from data. Data are split into training and evaluation datasets. A predictive algorithm classifying outcomes is fitted to the training dataset; its predictive power is assessed using the evaluation dataset.</td>
</tr>
<tr>
<td>Microbiological sampling criteria</td>
<td>EU regulations set out microbiological sampling criteria for key pathogen/animal pairings. Usually undertaken at processing (e.g. each swine abattoir must submit five samples per week for Salmonella testing).</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>Predictive analytics</td>
<td>The development of algorithms that can be used to categorise events and predict outcomes, commonly built using &quot;big data&quot; and machine learning methods</td>
</tr>
<tr>
<td>RFID</td>
<td>Radio Frequency Identification</td>
</tr>
<tr>
<td>RUSBoost</td>
<td>A decision tree machine learning algorithm used in binary classification cases where one classifier is more dominant than the other. (RUS – random under-sampling).</td>
</tr>
<tr>
<td>Track and trace</td>
<td>Track: record and monitor key epi and genetic information as food products flow through the food chain; trace: as for backtracing.</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Sequencing</td>
</tr>
</tbody>
</table>