UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA VETERINÁRIA

Comparison of Quantitative Microbial Risk Assessment Approaches Using *Listeria monocytogenes* in Serra da Estrela Cheese as a Case Study

Raquel de Lobo e Oliveira Costa

ORIENTADOR:
Doutor Patrick Njage

COORIENTADORES:
Doutor Maarten Nauta
Doutora Ana Rita Sá Henriques

2021

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA VETERINÁRIA

Comparison of Quantitative Microbial Risk Assessment Approaches Using *Listeria monocytogenes* in Serra da Estrela Cheese as a Case Study

Raquel de Lobo e Oliveira Costa

DISSERTAÇÃO DE MESTRADO INTERADO EM MEDICINA VETERINÁRIA

JÚRI
PRESIDENTE:
Doutora Maria João dos Ramos Fraqueza

VOGAIS:
Doutora Marília Catarina Leal Fazeres Ferreira
Doutor Patrick Njage

ORIENTADOR:
Doutor Patrick Njage

COORIENTADORES:
Doutor Maarten Nauta
Doutora Ana Rita Sá Henriques

2021

DECLARAÇÃO RELATIVA ÀS CONDIÇÕES DE REPRODUÇÃO DA DISSERTAÇÃO

Nome: Raquel de Lobo e Oliveira Costa

Título da Tese ou dissertação: Comparison of Quantitative Microbial Risk Assessment Approaches Using *Listeria monocytogenes* in Serra da Estrela Cheese as a Case Study

Ano de conclusão (indicar o da data da realização das provas públicas): 2021

Designação do curso de Mestrado ou de Doutoramento: Mestrado Integrado em Medicina Veterinária

Área científica em que melhor se enquadra (assinale uma):

☐ Clínica ☒ Produção Animal e Segurança Alimentar

☐ Morfologia e Função ☐ Sanidade Animal

Declaro sobre compromisso de honra que a tese ou dissertação agora entregue corresponde à que foi aprovada pelo júri constituído pela Faculdade de Medicina Veterinária da ULISBOA.

Declaro que concedo à Faculdade de Medicina Veterinária e aos seus agentes uma licença não-exclusiva para arquivar e tornar acessível, nomeadamente através do seu repositório institucional, nas condições abaixo indicadas, a minha tese ou dissertação, no todo ou em parte, em suporte digital.

Declaro que autorizo a Faculdade de Medicina Veterinária a arquivar mais de uma cópia da tese ou dissertação e a, sem alterar o seu conteúdo, converter o documento entregue, para qualquer formato de ficheiro, meio ou suporte, para efeitos de preservação e acesso.

Retenho todos os direitos de autor relativos à tese ou dissertação, e o direito de a usar em trabalhos futuros (como artigos ou livros).

Concordo que a minha tese ou dissertação seja colocada no repositório da Faculdade de Medicina Veterinária com o seguinte estatuto (assinale um):

1. ☒ Disponibilização imediata do conjunto do trabalho para acesso mundial;
2. ☐ Disponibilização do conjunto do trabalho para acesso exclusivo na Faculdade de Medicina Veterinária durante o período de ☐ 6 meses, ☐ 12 meses, sendo que após o tempo assinalado autorizo o acesso mundial*;

* Indique o motivo do embargo (OBRIGATÓRIO)

Nos exemplares das dissertações de mestrado ou teses de doutoramento entregues para a prestação de provas na Universidade e dos quais é obrigatoriamente enviado um exemplar para depósito na Biblioteca da Faculdade de Medicina Veterinária da Universidade de Lisboa deve constar uma das seguintes declarações (incluir apenas uma das três):

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.
2. É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA TESE/TRABALHO (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.) APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.
3. DE ACORDO COM A LEGISLAÇÃO EM VIGOR, (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.) NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA TESE/TRABALHO.

Faculdade de Medicina Veterinária da Universidade de Lisboa, 19 de Outubro de 2021

(indicar aqui a data da realização das provas públicas)

Assinatura: Raquel Costa

Acknowledgements

To professor Telmo Nunes for introducing me to the world of epidemiology and risk analysis, for helping me plan my internship and for guiding me to obtain the skills needed to follow a career in this field.

To my supervisor Parick Njage and co-supervisor Maarten Nauta for the challenges proposed, teachings and support given during my internship at DTU. Also, to Clementine Henri who taught me the basic concepts of bioinformatics and helped me perform the bioinformatic analyses which were crucial for this study.

To my co-supervisor Ana Rita Henriques for her guidance prior and during my internship, and for the contributions that helped me write this dissertation.

To the friends I met in Denmark for making me feel welcome and for making this experience unforgettable.

To my friends Carolina, Bruna, Rita and Patrícia with who I shared these last 6 years. For all the lunches, group works, phone calls to share the last thoughts and questions before the exams, and for all the adventures we have been through together.

To my long-standing friends who have always been there for me.

To Paul for the endless support and encouragement, for making everything easier and less stressful.

To my parents, my sister, and my family for always believing in me and for the unconditional support.

Comparação de Abordagens Quantitativas de Avaliação de Risco Microbiológico Usando *Listeria monocytogenes* em Queijo Serra da Estrela como Estudo de Caso

Resumo

O uso de metodologias de sequenciação total do genoma (WGS) bacteriano tem vindo a aumentar nos últimos anos, permitindo a sua aplicação em Avaliação de Risco Microbológico (AQRM). Neste estudo procurou-se comparar a AQRM clássica com a AQRM baseada em WGS, de modo a determinar as questões da avaliação de risco a que conseguem responder e o apoio que fornecem na tomada de decisão da gestão do risco. Para isso, utilizou-se a *Listeria monocytogenes* em queijo Serra da Estrela como estudo de caso.

Realizou-se inicialmente uma revisão bibliográfica sobre AQRM clássica e a aplicação de WGS na AQRM. Esta revisão evidenciou que a integração dos dados resultantes da metodologia de WGS em AQRM clássica é vantajosa nos vários passos da AQRM. Após a revisão bibliográfica, foi desenvolvida uma AQRM clássica que estimou um total de 16 casos de listeriose, em Portugal, num ano, devido ao consumo de queijo Serra da Estrela. Múltiplos cenários alternativos foram testados e mostraram a importância da conservação do queijo Serra da Estrela em refrigeração durante a vida útil. A AQRM baseada em dados resultantes de WGS, utilizou um modelo de *machine learning* modelado com dados franceses obtidos através de WGS provenientes de estirpes de *L. monocytogenes*, cuja frequência em casos clínicos humanos era conhecida. Inserindo dados de WGS de *L. monocytogenes* isolada de múltiplos queijos, o modelo previu uma frequência de casos de listeriose entre 37 e 54% devido ao consumo de queijo Serra da Estrela. Este modelo identificou ainda os genes de *L. monocytogenes* e os queijos mais frequentemente associados a casos clínicos de listeriose.

O estudo concluiu que as AQRM respondem a questões diferentes e apoiam diferentes medidas de controlo. O AQRM clássico fornece a informação científica necessária para os gestores de risco decidirem quais as melhores estratégias de mitigação do risco, enquanto o AQRM baseado em WGS permite a rápida deteção de surtos e uma tomada de decisão mais informada na retirada de produtos da cadeia alimentar. Este estudo sugere que a construção de modelos para prever a frequência de casos clínicos é útil para os gestores de risco, uma vez que os dados de WGS podem ser integrados na AQRM clássica para obter estimativas mais precisas e ser usados independentemente, como ferramenta de atuação imediata. Mais estudos relativos ao uso de dados resultantes de WGS em AQRM para a tomada de decisão na gestão de risco são necessários para garantir um uso adequado da informação.

Palavras-chave: avaliação do risco, whole genome sequencing, machine learning, *Listeria monocytogenes*, queijo

Comparison of Quantitative Microbial Risk Assessment Approaches Using *Listeria monocytogenes* in Serra da Estrela Cheese as a Case Study

Abstract

Whole Genome Sequencing (WGS) data has been growing and its recent application in Quantitative Microbial Risk Assessment (QMRA) has been discussed. Taking this growth into consideration, this study compared classic QMRA and WGS QMRA in terms of risk assessment questions they can answer and the way they can support decision making by risk managers. For this purpose, *Listeria monocytogenes* in Serra da Estrela cheese was used as a case study.

Initially, a review of existing literature related to classic QMRAs and the application of WGS in QMRA was performed. This review revealed that WGS has shown advantages when integrated in the classic QMRA by allowing to fine tune each step of the risk assessment.

After literature review, a classic QMRA was performed which predicted a total of 16 listeriosis cases in Portugal in one year due to the consumption of Serra da Estrela cheese. Multiple scenarios were tested, and results underline the importance of the cheese being stored at refrigeration temperatures. The WGS QMRA based on available WGS data of *L. monocytogenes* isolated from cheeses, using a machine learning model trained with French *L. monocytogenes* WGS data with known clinical frequency, predicted a clinical frequency of 37 to 54% due to Serra da Estrela cheese consumption and identified the genes and cheeses that are associated the most with clinical cases.

This study concluded that both the assessed QMRA approaches are good in answering different questions and may support different types of control measures. Classic QMRA is good in giving the necessary scientific information for risk managers to decide on mitigation strategies whereas WGS QMRA allows for an early detection of outbreaks and more informed decision on product withdrawal. Therefore, this study suggests that having models to predict the clinical frequency based on WGS can be useful for risk managers as WGS data can, not only be integrated in the classic QMRA to obtain more precise results, but also be used independently as a first approach tool to promptly detect outbreaks and decide if immediate measures are required. However, further studies on the use of WGS for decision making in the risk management phase are needed for a correct use of the information.

Keywords: risk assessment, whole genome sequencing, machine learning, *Listeria monocytogenes*, cheese

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMR - Antimicrobial Resistance
BLAST - A Basic Local Alignment Search Tool
bp - Base Pairs
CFU - *Colony Forming Units*
cgMLST - Core Genome Multilocus Sequencing Typing
DNA - Deoxyribonucleic Acid
DTU - Technical University Of Denmark
ECDC - European Centre For Disease Prevention And Control
EEA - European Economic Area
EFSA - *European Food Safety Authority*
EGR - Exponential Growth Rate
ENA - European Nucleotide Archive
EU - European Union
g - *Gram*
gQMRA - Generic Quantitative Microbial Risk Assessment
GWAS - Genome-Wide Association Studies
LB - *Logit Boost*
MLST - *Multilocus Sequencin Typing*
NCBI - National Center For Biotechnology Information
NGS - Next Generation Sequencing
NIR - No Information Rate
PCR - Polymerase Chain Reaction
QMRA - *Quantitative Microbial Risk Assessment*
RF - *Random Forest*
RMSE - Root Mean Squared Error
RTE - *Ready-To-Eat*
SVM - *Support Vector Machine*
TEO - *Total Number Of Eating Occasions Per Year*
WGS - *Whole Genome Sequencing*

1. Internship Report

From September to December 2020, an internship under the supervision of professor Telmo Nunes took place at the Faculdade de Medicina Veterinária, Universidade de Lisboa. During this period, data analysis skills focused on epidemiological analysis using R programming language were acquired and more advanced epidemiology concepts were learned and applied through the development of three projects.

The main project focused on exploring the correlations between covid-19 prevalence and population mobility. The study aimed to understand whether measures such as confinement and limiting population movement were significantly effective in controlling the disease, considering that their application has a major impact on the life of the population and the country's economy. For this project, Facebook and Google databases were used to analyse citizens' mobility, and Data Science for Social Good Portugal databases were used to analyse covid-19 cases, hospitalisations, and tests. After data curation and analysis, results indicated that population mobility had different impacts on disease prevalence depending on the implementation of mitigation strategies and those impacts were verified within seven to nine days. The final reports and a web app developed using Shiny App are available in the sharing platform Github (https://github.com/Raquel-Costa/epi_intern_fmv-ul). During in person and online meetings, critical thinking for problem solving and results presentation skills were trained.

In addition, a project regarding animal health data of swine production facilities in Portugal owned by Portuguese Veterinary Services was performed. Using Structured Query Language, prevalence of Aujeszky disease, vaccination patterns, and other factors were analysed. Additionally, a study looking at potential infection of animals whose owners have covid-19 aimed to explore pet species potentially affected by the disease, and potential transmission pathways. The study population and research questions were defined, questionnaires were developed, and database structure was defined.

To complement this knowledge, the Statistics with R course was taken (available at the Coursera platform: https://www.coursera.org/specializations/statistics#courses). In order to acquire further knowledge regarding future internship topics, metagenomic course (available at the Coursera platform: https://www.coursera.org/learn/metagenomics), whole genome sequencing course (available at the Coursera platform: https://www.coursera.org/learn/wgs-bacteria) and the theoretical part of the machine learning course (available at the Coursera platform: https://www.coursera.org/learn/machine-learning) were taken.

From January to May 2021, the internship continued with the supervision of Professor Patrick Njage and the co-supervision of Professor Maarten Nauta and Professor Ana Rita Sá Henriques at the National Food Institute, Technical University of Denmark. The aim of the

internship was to compare two quantitative microbial risk assessment approaches. To achieve this goal, the internship started with a literature review which promoted the acquisition of knowledge regarding quantitative microbial risk assessment, why and how it is done, obtained outputs and potential applications of whole genome sequencing in the process. Then this knowledge was put into practice by performing a classic quantitative microbial risk assessment, as well as a quantitative microbial risk assessment with whole genome sequencing data. During this process, skills related to Excel, R programming language, data preparation, data analysis, data visualization and interpretation of results were developed. The integration of a multi-disciplinary team allowed the opportunity to learn about bioinformatics and machine learning and to obtain hands-on experience. Critical thinking and the follow-ups including presentations of the developed work, were crucial to reach the final results of the study.

2. Literature Review

2.1. Quantitative Microbial Risk Assessment

Risk assessment is the scientific component of risk analysis. It is defined as "characterizing the potential adverse effects to life and health resulting from exposure to hazards over a specified time period" (EDES 2012). Risk managers, in the risk management component of the risk analysis, decide when risk assessment is needed. Complex hazard exposure pathway, incomplete data on the hazard or clinical outcome, significant regulatory or stakeholder concerns, mandatory regulatory requirement for a risk assessment and the need to verify that a response to an urgent food safety problem is scientifically justified, are the main reasons to perform risk assessment (EDES 2012).

One type of risk assessment is the Quantitative Microbial Risk Assessment (QMRA) in which the hazard involved is a pathogen. It aims to assess the consumers' risk of illness due to a pathogen present in food and allows the evaluation of proposed intervention measures aiming to mitigate public health risk which is very useful for risk managers (Nauta 2008). As its name indicates, QMRA is a quantitative method, meaning that the outputs are expressed numerically and may include a numerical description of uncertainty. In a quantitative risk assessment two approaches can be considered. The first approach is deterministic, also called point estimate, consisting of a single numerical value representing for example the risk in a worst-case scenario or an average risk (more common for chemical risk assessment). The second approach is stochastic, also called probabilistic risk estimates, which includes variability and uncertainty and is presented as a distribution reflecting more real-life situations, but it is often complex and difficult to generate requiring mathematical modeling of the

variability of the phenomena involved (more common for QMRA) (EDES 2012). Therefore, QMRA is a quantitative method, normally with a stochastic approach.

The first step of QMRA is hazard identification which is a qualitative evaluation of available knowledge, compiling important information on the pathogen, food product and host interface. Epidemiological investigations, source attribution, national surveillance databases, microbial research, process evaluations and clinical studies should be analysed to understand where the problem is and its extend (Fazil 2005).

QMRA second step is hazard characterization which involves the use of existing data and literature information to develop a dose-response model. The dose-response model describes the probability of a specified response from exposure to a certain pathogen in a population, as a function of the dose. It links the dose of the pathogen ingested with the response by estimating the probability of illness upon exposure to the hazard which must account for the probability of the pathogen being ingested, surviving, and infecting the host and the probability of disease development after infection (FAO and WHO 2003; Fazil 2005).

QMRA third step is exposure assessment which characterizes the amount of pathogen that is consumed by each exposed population. For this purpose, levels of hazard in raw materials, in food ingredients and in the food environment are used to track changes in levels throughout the food production chain. These data are combined with food consumption patterns of the target consumer population to assess exposure to the hazard over a particular period of time (EDES 2012; EFSA Panel on Biological Hazards et al. 2019). As detailed exposure data characterizing the extent of microbiological hazard present in foods at the time of consumption are usually not available, exposure assessment will commonly require the development of mathematical models in which all relationships between factors affecting exposure can be described mathematically and using logical tests and conditional statements in the model. In an exposure assessment, input variables would include factors such as time, temperature, production volume and dilution during processing (FAO and WHO 2008).

The fourth and last step of QMRA is risk characterization through which the probability of illness, called risk, is estimated. This probability will be used by risk managers to assist their decision-making process. In this step, outputs from the previous three steps are integrated to generate an estimate of risk. Risk estimates can be calculated in terms of risk per serving, per day, per year, number of illnesses per year, or some other similar variation. A risk characterization often includes narrative on other aspects of the risk assessment, such as comparative rankings with risks from other foods, impacts on risk of various “what if” scenarios, and further scientific work needed to reduce data and knowledge gaps. Besides risk estimation, uncertainty and variability must also be described if possible (EDES 2012; EFSA Panel on Biological Hazards 2019). Uncertainty is the lack of perfect knowledge of parameters which can be reduced by measurements whereas variability is the true heterogeneity of

population due to the physical system and cannot be reduced by measurements (it can be variability in time, space, population, product, and others) (Nauta 2008). Variability is a characteristic of phenomena that differ from one observation to the next, for example, people eat different amounts of a food, and the level of a particular hazard present in a food can also vary from one serving of food to another (EDES 2012). Variability is important in risk assessment because final population risk is given by the mean risk as calculated from a distribution describing the variability of ingested doses (Bassett et al. 2012).

2.2. Serra da Estrela cheese

Serra da Estrela cheese is a Portuguese cured semi-soft cheese obtained from raw milk of Bordaleira Serra da Estrela and/or Churra Mondegueira sheep breeds. It weights between 0.5 to 1.7 kg and during its production salt and coagulant are added to the sheep milk that will further go through coagulation, curd work, pressing, external salting and ripening for at least 30 days and no more than 120 days. This cheese has a Protected Designation of Origin and is one of the most appreciated and one the most important in terms of economy in Portugal (Guilherme 2012).

2.3. Listeria monocytogenes

Recognized as pathogenic for animals in 1927 and a threat to public health in the 80s after food consumption that led to severe outbreaks, *Listeria monocytogenes* is a potentially pathogenic bacteria that causes listeriosis (ASAE [date unkown]).

Human non-invasive listeriosis is a mild form of the disease that leads to febrile gastroenteritis, whereas human invasive listeriosis is a systemic, life-threatening disease that usually affects individuals with underlying conditions that impair their immune response (Pouillot et al. 2009a). Pregnant women, newborn, elderly and immuno-compromised patients, are considered risk groups (Huang and Hwang 2012). Immunocompromised adults usually experience septicaemia and meningitis, while pregnant women often present nonspecific symptoms such as fever and prostration, followed by abortion, stillbirth, premature birth or newly born with bacteraemia and meningitis (Tirloni et al. 2018). For these vulnerable populations, the risk of fatal listeriosis is 10 to 100 times higher than for the rest of the population (Campagnollo et al. 2018a). Normally, only invasive listeriosis is diagnosed as people with non-invasive listeriosis may not seek medical help, and/or are prescribed a treatment without bacterial isolation and identification.

The European Centre for Disease Prevention and Control (ECDC) collects data on infectious diseases being listeriosis one of them (ECDC [date unknown]). For a listeriosis case

to be reported, patients need to present one of the following five symptoms: fever, meningitis/meningoencephalitis/encephalitis, influenza-like symptoms, septicemia or localized infections such as arthritis, endocarditis, endophthalmitis and abscesses. If listeriosis involves a pregnant woman miscarriage, stillbirth or premature birth are possible consequences. Listeriosis is reported in newborns if there is stillbirth, or premature birth, or one of the following five symptoms: meningitis/meningoencephalitis, septicemia, dyspnea, granulomatosis infantiseptica or lesion on skin, mucosal membranes or conjunctivae (ECDC [date unknown]).

The most recent data indicates 2652 listeriosis confirmed cases in European Union (EU) and European Economic Area (EEA) in 2019 (**Figure 1**) (ECDC [date unknown]). The most affected were those over 64 years of age (ECDC [date unknown]). In Portugal, listeriosis became a mandatory notifiable disease in April 2014. Until 2018, the yearly number of cases was rising, reaching 68 cases in 2018. However, in 2019 a decrease was found and 56 listeriosis confirmed cases were notified (**Figure 2**) (ECDC [date unknown]).

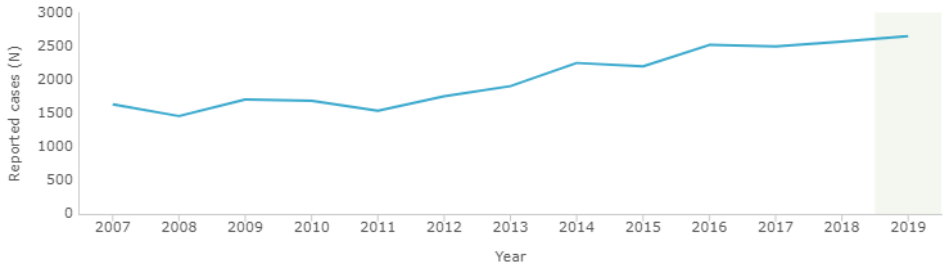

Figure 1 - Listeriosis reported cases in EU/EEA (ECDC [date unknown]).

Figure 2 - Listeriosis reported cases in Portugal (ECDC [date unknown]).

In 2019, the fatality rate considering all reported cases with known outcome was 17.5% in EU/EEA and 18.2% in Portugal (**Figure 3**). Also in 2019, the hospitalization proportion reached 92.2% in EU/EEA and 98.2% in Portugal (**Figure 4**). These rates demonstrate the high frequency of deaths and hospitalizations due to listeriosis, justifying its importance in public health, despite its low prevalence (Njage et al. 2019; ECDC [date unknown]).

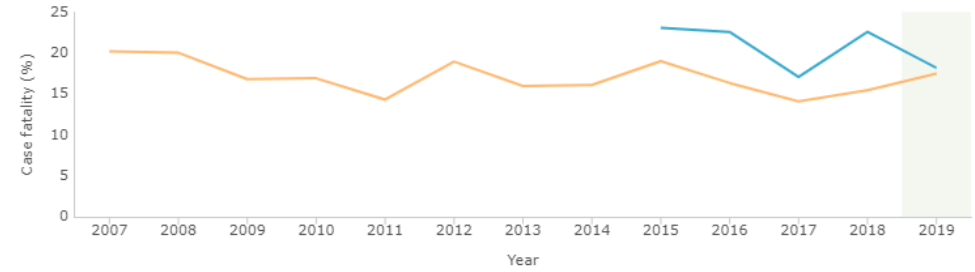

Figure 3 - Listeriosis case fatality rate in EU/EEA (yellow) and Portugal (blue) (ECDC [date unknown]).

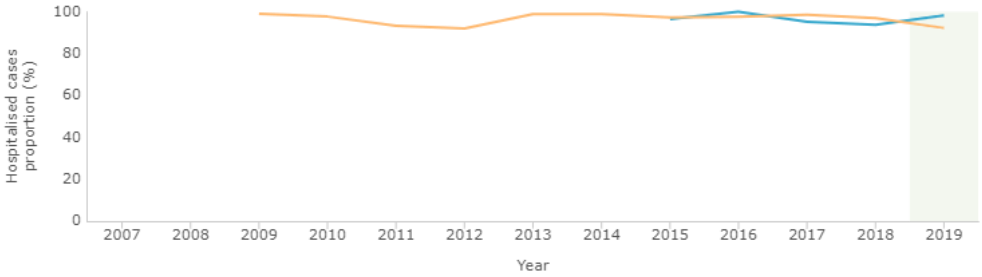

Figure 4 - Listeriosis hospitalised cases proportion in EU/EEA (yellow) and Portugal (blue) (ECDC [date unknown]).

L. monocytogenes has an ubiquitous distribution and, as animals and humans can be asymptomatic carriers, contamination of raw material and foods is frequent, especially food products that present good conditions for bacterial growth and long shelf-life (ASAE [date unknown]). A wide variety of raw and processed foods, including milk and dairy, meat, eggs, and their derived products, as well as seafood and vegetables, may be contaminated with *L. monocytogenes* (Tirloni et al. 2018). Contaminated food is estimated to be the source of as many as 99% of listeriosis cases (Njage et al. 2018).

Quantitative modelling suggests that more than 90% of invasive listeriosis is caused by ingestion of ready-to-eat (RTE) food containing more than 2000 colony forming units (CFU) per gram (g), and that one-third of cases are due to bacterial growth in the consumer phase. RTE foods are a group of food products that are pre-cleaned, precooked, mostly packaged and ready for consumption without prior preparation or cooking. *L. monocytogenes* is the major concern in refrigerated RTE foods (Huang and Hwang 2012). With this major concern, food safety criteria for *L. monocytogenes* in RTE foods have been established and applied by food business operators from 2006 onwards (EFSA Panel on Biological Hazards et al. 2018a)

Based on quantitative risk characterisation of *L. monocytogenes* in various RTE food categories, the food subcategory associated with the largest number of cases per year was cooked meat, followed by sausage, gravad fish, cold-smoked fish, pâté, soft and semi-soft cheese and hot-smoked fish (EFSA Panel on Biological Hazards et al. 2018a).

2.4. Classic QMRA Model for *L. monocytogenes*

A review through existing studies of classic QMRA on *L. monocytogenes* was made and a summary of each study is present in Appendix 1. These studies use multiple approaches and equations to assess the probability of illness due to the ingestion of *L. monocytogenes* in a specific food product. However, when it comes to *L. monocytogenes* in of RTE foods, the European Food Safety Authority (EFSA) recently developed a generic quantitative microbial risk assessment model (gQMRA) for *L. monocytogenes* contamination of RTE foods. *L. monocytogenes* has the ability to survive and develop during refrigerated storage, increasing the risk of illness when compared to other bacteria that have their growth decelerated by low temperatures (Njage et al. 2020).

The gQMRA model predicts consumer exposure to *L. monocytogenes* based on the initial contamination level at retail of a variety of RTE foods, and the potential growth before consumption. The probability of a consumer being infected and developing listeriosis is then predicted by applying a dose-response model (EFSA Panel on Biological Hazards et al. 2018a). The gQMRA was developed in R (R Core Team 2020, Vienna, Austria) and allows an expanded evaluation of uncertainty when the uncertainty about the inputs is available. The model also allows inclusion of the variability related to exposure assessment (EFSA Panel on Biological Hazards et al. 2018b). This model is called generic because users can add food categories and their own data, aiming to estimate the risk associated with consumption of different RTE food categories for all age and sex groups. It is important to notice that the model accepts multiple food products as an input and does not distinguish risk associated with each food product, only the risk associated with all the food products combined (EFSA Panel on Biological Hazards et al. 2018a).

2.5. Whole Genome Sequencing and Quantitative Microbial Risk Assessment

2.5.1. Whole Genome Sequencing and Bioinformatics

The genome is the entire genetic material of a living organism, such as bacteria and eukaryotes, consisting of deoxyribonucleic acid (DNA). Each organism has a unique DNA sequence that is composed of a combination of four different nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). The unique DNA of an organism is identified by knowing the sequence of the bases. Determining the order of bases is called sequencing. Whole genome sequencing (WGS) is a laboratory procedure that determines the order of bases in the genome of an organism in one process (CDC 2016).

The process begins with the DNA isolation. Nowadays next generation sequencing (NGS) technologies cannot sequence fragments longer than 100 base pairs (bp). Therefore,

a library preparation step is needed. This step allows preparation of the genetic material into a form that is compatible with the sequencing instrument. This can include fragmentation, end repair, cleaning of the DNA, polymerase chain reaction (PCR) step to increase the amount of DNA, and adapter ligation. After library preparation, DNA is sequenced using most of the time Illumina platform. The output of the Illumina sequencing consists of reads that are continuous sequences of around 100 bp of DNA produced when a sequencing instrument analysed the genetic material. All the reads are gathered into a FASTq file which contains the DNA sequences as well as quality information for each base (Duarte et al. 2018, Ekblom and Wolf 2014).

Then, a DNA sequence analysis is conducted from the reads in the FASTq file. At this point, bioinformatics, the use of computers programs to acquire, manage and analyse biological information, is required. The first step of bioinformatics analysis is the assembly. The assembly is the process by which all the reads are put together attempting to reconstruct the original DNA sequence. This step is needed because, as mentioned before, current sequencing technology cannot read whole genomes as a continuous sequence, but rather short fragments (reads). When reads are merged, they create contigs which are stretches of continuous, inferred DNA sequence, resulting from assembling shorter sequencing reads. There are two types of assembly: *de novo* assembly where the contigs are inferred using assembly algorithm, and mapping assembly where contigs are formed based on a reference sequence from the same species. When the assembly is completed, the quality of the contigs has to be verified. Quality control includes two major parameters the N50 and the number of contigs. The N50 is a statistical measure of the length of contigs. It is the median of the lengths of contigs and the longer it is, the better is the assembly. The other parameter is the number of contigs, in which fewer contigs are better, as it means that the contigs are longer, with no gaps and complete sequence. For this last parameter there is a breakpoint where a sample with more than 500 contigs is considered to have low-quality (Cavaco and Leekitcharoenphon 2017; Duarte et al. 2018, Ekblom and Wolf 2014).

After assembly, multiple analysis can be conducted such as Multi-Locus Sequencing Typing (MLST) using seven housekeeping genes, in which each sample is given a number to each gene according to the allele present and this combination is called a sequence type; core genome Multilocus Sequencing Typing (cgMLST) that instead of using seven housekeeping genes, uses core genes, present in all strains of a species; and pangenome. Pangenome is an analysis that obtains the core genome and accessory genome. The variable or accessory genome refers to genes that are not present in all strains of a species (Scholz [date unknown]). The pangenome can be obtained using two steps process. The first involves analysis using Prokka, which is a rapid prokaryotic genome annotation that allows the identification of relevant parts of the genome and its labelling. It uses multiple subprogrammes, in particular the

Prodigal, a tool that detects and translates genes into proteins (Seemann 2014). The second step involves the use of Roary, a pipeline that uses the Prokka output to obtain the pangenome from the collection of *L. monocytogenes* isolates (Page et al. 2015).

Finally, after the sequence analysis made to obtain results, a statistical analysis is performed to interpret results. For this purpose, a set of data that describes and gives information about other data, called metadata, is frequently used (Duarte et al. 2018, Ekblom and Wolf 2014).

2.5.2. WGS and Machine Learning

With such an increase in the size of datasets, two solutions arise to analyse such a big amount of data, avoiding irrelevant biological outcomes and the discharge of important data. The first one is the use of network-based analysis techniques and the second one is the use of machine learning algorithms. For the second one, the application of computer algorithms that improve with experience will allow identification of predictor combinations that will predict the risk outcome, turning big data sets to fewer number of predictors. With machine learning it is possible to reveal the properties of a sequence that are most important for determining a certain phenotype or to predict the occurrence of a protein (Njage et al. 2018).

Machine learning is a branch of artificial intelligence that focuses on data and algorithms to imitate the way humans learn and make classifications or predictions based on that knowledge. It was defined in 1959 by Arthur Samuel as “the field of study that gives computers the ability to learn without being explicitly programmed”. There are three types of machine learning, i) supervised learning in which the algorithm makes predictions based on a set of examples, ii) unsupervised learning in which the algorithm makes predictions based on unlabelled data, and iii) semi-supervised learning where a smaller labelled dataset is used to guide classification and a larger unlabelled data set is used for feature extraction (Delua 2021).

To analyse WGS data, supervised learning is most frequently used. This type of learning is divided into classification, predicting a discrete valued output, and regression, predicting a continuous valued output (Delua 2021).

Single models can be used for both classification and regression. One example is the support vector machine (SVM), a supervised learning algorithm used for both classification and regression. This algorithm creates a hyperplane that divides the data based on the features allowing predictions. SVM takes advantage of the kernel trick to use a linear classifier to solve non-linear problems. There are multiple types of kernels such as linear kernel, polynomial kernel and radial kernel (Ray 2017). However, ensemble methods have shown better results than single models. Ensemble methods create multiple models and combine them together to obtain more accurate results. The ensemble methods can be classified as:

sequential methods in which there is dependency on previous data, parallel methods where the dependency does not exist, homogeneous ensemble using the same classifier but with different datasets and heterogeneous ensemble using different classifiers but with the same datasets (Pedamkar 2020).

One of the most used ensemble methods is bagging, a parallel method based on the boostrapping principle meaning that samples are created with replacement from the original data and the aggregating principle where the final output is the most common result (also called voting when used in classification) or the average (for regression) of the results obtained in each model (Pedamkar 2020). Random Forest (RF) is an example of a machine learning algorithm that uses bagging. RF, a supervised learning algorithm used for both classification and regression, is one of the most used machine learning algorithms. It builds multiple decision trees based on different features and each decision tree makes a prediction for the outcome. The most common prediction among all the decision trees' predictions is the final prediction. Overall, this algorithm has low probability of overfitting (when the algorithm fits the training set very well but fails to generalise and predict on new examples), has high accuracy and estimates missing data, however it is relatively slow, and its use is limited in regression problems (Donges 2021).

The other very common ensemble method is boosting, a sequential method that places weak learner sequentially and each model will try to correct the error of the previous model (Pedamkar 2020). Logit Boost (LB) is an example of a boosting classification algorithm (Sun et al. 2014).

To evaluate the machine learning model performance, some metrics are used. Accuracy and kappa are the most common ones for classification models. Accuracy is the number of outcomes predicted correctly out of all outcomes, expressed from 0 to 1 where 1 indicates all outcomes predicted correctly. Kappa, or Cohen's kappa, is an accuracy normalized at the baseline of random chance on the dataset, ranging from -1 to 1 where values lower than 0 implies no agreement between the observed and predicted classes and 1 suggests perfect concurrence between the predicted and observed classes. Root Mean Squared Error (RMSE) and R Squared are the most common metrics for regression models. RMSE is the average deviation of the predicted outcomes from the real outcomes while R squared indicates the proportion of the variance in the outcome that is predictable by the features (Brownlee 2019).

2.5.3. WGS Application in QMRA

WGS can detect single nucleotide variants, insertions, deletions, and large structural variants and its benefits can be applied to each step of QMRA (Ronholm et al. 2016). Multiple

studies, summarized in Appendix 1, have analysed the possible applications of WGS data in classic QMRA.

In hazard identification step, WGS allows the identification of the hazard at the strain level and the detection of variability in virulence profiles among strains. This information promotes the development of a finer QMRA, targeting the epidemiologically relevant pathogen–food combinations (EFSA Panel of Biological Hazard et al. 2019). One example is Njage et al. (2018) study, identifying food products in which genes associated with high illness frequencies were present. WGS can also allow the detection of casual links of exposure by comparing genomic profiles between human clinical isolates and isolates collected along the exposure pathway (EFSA Panel of Biological Hazard et al. 2019). Sunde et al. (2015) found similar multiresistance plasmids in pork meat and in humans infected with *Escherichia coli,* indicating meat as a possible source of antimicrobial resistance (AMR). Day et al. (2017) also found similar AMR profile in *E. coli* from ruminants and from humans indicating possible transmission of AMR *E. coli* from animals or their environment to humans.

Regarding hazard characterization, WGS data giving information on variability in virulence profiles among strains along with phenotypic data can lead to the creation of better dose-response models. These models will allow more targeted risk assessment to pathogen–human interactions, since virulent strains may have a higher probability of causing foodborne infection and different strains may cause different clinical outcomes. Phenotypic data is very important, as the expression of virulence genes may be affected by several environmental conditions along the food chain and after consumption (EFSA Panel of Biological Hazard et al. 2019). Fritsch et al. (2018) defined three classes of *L.* monocytogenes according to their virulence potential and associated each class with a different dose–response model.

In exposure assessment, WGS can be used to identify and track useful markers for predicting microbial behaviour in foods, often through genome-wide association studies (GWAS) which compare a large set of genomic data and associate them to specific phenotypic traits, allowing for the identification of genomic sequences as markers or indicators of specific phenotypes. WGS can be used to predict the ability of a microorganism to grow or survive within the host or the food, as well as during processing, storage and distribution of foods. Fritsch et al. (2019) conducted a GWAS where a number of genes were identified as associated with *L. monocytogenes* growth at low temperature (2°C). In addition, Njage et al. (2020) used WGS data with known stress response to acid, cold, salt and desiccation to predict stress response in new WGS data from samples where the stress response is unknown, through machine learning predictive models.

As WGS data can be applied in each step of the risk assessment, risk characterisation also benefits from this type of data, as this step is a combination of the previous ones (EFSA Panel of Biological Hazard et al. 2019). In this matter, Fritsch et al. (2018) illustrated the

potential of WGS to refine QMRA by updating the number of listeriosis cases associated to cold-smoked salmon in France with information on the characteristics of different *L. monocytogenes* clonal complexes, including the identification of genetic markers for the ability of strains to grow at low temperatures and for their virulence potential. Results showed that uncommon highly virulent strains and strains with a low minimal growth temperature were responsible for the majority of predicted human cases. Njage et al. (2020), as mentioned before, presented an example of QMRA using strains of unknown stress phenotype. Increased resistance to stress conditions leads to increased growth, likelihood of higher exposure and probability of illness. Neglecting within-species genetic and phenotypic heterogeneity in microbial stress response may over or underestimate microbial exposure and eventual risk during QMRA.

This new way of QMRA allows focusing on a strain, especially those that are more virulent and more important from a public health standpoint. It also allows to better explore dose-response models when combined with phenotype data and to predict the microbial behaviour in host, food, processing, storage and distribution as biomarkers responsible for variability regarding growth and survival of the microorganism can be detected. Finally, it allows comparison of the microorganism genome obtained from a human with the microorganism genome obtained from each step of the food production which can help identify casual links of exposure (EFSA Panel of Biological Hazard et al. 2019).

WGS permits dealing with strain variability which allows to fine tune the hazard identification, dose-response, and exposure models in QMRA. It can also allow risk managers to prioritise hazards more accurately in risk ranking. However, it still has some limitations hence the need for phenotypic data (EFSA Panel of Biological Hazard et al. 2019).

2.6. WGS and *L. monocytogenes*

As mentioned before and detailed in Appendix 1, some WGS studies were made in *L. monocytogenes.* EFSA scientific opinion on *L. monocytogenes* contamination of RTE foods and the risk for human health in the EU, recommends innovative methodologies including WGS for strain identification and monitoring of trends. WGS techniques, when combined with epidemiological information, have the potential to attribute relatedness among *L. monocytogenes* strains and thus establish stronger links between human listeriosis cases and causative foods (EFSA Panel on Biological Hazards et al. 2018a).

3. Objectives

The main goal of this study was to compare classic QMRA and WGS QMRA in terms of risk assessment questions they can answer and the way they can support decision making by risk managers. This was done on the basis of a review of existing studies, summarized in Appendix 1, and a case study on *L. monocytogenes* in Serra da Estrela cheese. This case study was selected as *L. monocytogenes* is a potentially pathogenic bacteria for which the importance of WGS has been identified, and Serra da Estrela cheese is a RTE food product, in which *L. monocytogenes* is a concern, with existing data regarding the cheese and its consumption.

The classic QMRA assessed the probability of listeriosis and the expected number of listeriosis cases due to *L. monocytogenes* in Serra da Estrela cheese for each population group, defined by age range and sex, using the gQMRA method mentioned in section 2.4. and available data on the cheese. For WGS QMRA there was a database available associating strains to their clinical frequency. Therefore, this QMRA predicted the clinical frequency of *L. monocytogenes* in Serra da Estrela cheese using bioinformatics and machine learning.

In section 4.1., data collection and preparation are detailed. Firstly, the origin of data for classic QMRA is detailed, explaining the assumptions and calculations to obtain some required parameters. Then the source of data for WGS QMRA are mentioned, explaining the bioinformatic tools used to obtain data that can be used for QMRA. In section 4.2., the data analysis for classic QMRA are mentioned, exploring all the steps of the gQMRA, followed by data analysis for WGS QMRA, using machine learning. Finally, results are presented in section 5. for each QMRA and a discussion is made in section 6. comparing both QMRAs and answering the questions of this study.

4. Material and Methods

4.1. Data collection and preparation

This section is divided in classic QMRA and WGS QMRA and describes data and transformations applied to perform both QMRAs of *L. monocytogenes* in Serra da Estrela cheese.

4.1.1. Classical QMRA

To perform the classic QMRA, data on retail and consumption of Serra da Estrela cheese was obtained from Guilherme (2012). A summary of the important parameters is presented in **Table 1**.

Table 1 - Overview of variables and parameters obtained from Guilherme (2012) study on Serra da Estrela cheese.

Variable/Parameter	Value	Source
Prevalence of *L.monocytogens* (P)	39.6%	Analysis of 91 Serra da Estrela cheeses from retail (Manuela Sol, Personal Communication to Guilherme on 5 May 2011)
Domestic storage temperatures (T)	Min = -0.5ºC Mean = 17ºC Max = 34.5ºC	Record of cheese temperatures every 15 minutes in a sample of 39 Serra da Estrela cheeses given to consumers
Domestic storage time (t)	Min = 1 day Most likely = 9 days Max = 21 days	Questionnaires made to 39 Serra da Estrela cheeses family aggregates accounting for a total of 107 consumers
Serving size (wt)	Mean = 44.1 g SD = 56.7 g Min = 15 g Most likely = 20 g Max = 60 g	
Consumer age and susceptibilities' profile	≤ 4 years = 4% 5-14 years = 9% 15-59 years = 64% ≥ 60 years = 21% Pregnant = 1% Immunocompromised = 1%	
Weight of the cheese (wt_{cheese})	0.5 to 1.7kg Mode = 0.5 kg	Based on Guilherme (2012) statement and in an online search through the biggest supermarket chains in Portugal
Consumptions per year (c)	3 occasions (Christmas, Easter and summer holidays)	Questionnaires made to 39 Serra da Estrela cheeses family aggregates accounting for a total of 107 consumers and survey to the products with a protected designation of origin management consortium
Number of people eating one cheese	Mean = 3 SD = 2	Questionnaires made to 39 Serra da Estrela cheeses family aggregates accounting for a total of 107 consumers

The parameters prevalence (P), domestic storage temperatures (T), domestic storage time (t), serving size (wt), weight of the cheese (wt_{cheese}) and consumptions per year (c) described in **Table 1** were used to build gQMRA. However, some parameters needed for the analysis required transformation and/or were not presented in Guilherme (2012) study. These parameters and the assumptions made to obtain them are described in the following sections.

4.1.1.1. Serving size

The serving size (wt) is not given by age and sex. As this differentiation is made on the gQMRA model and as it was considered important for the risk assessment, adjustments of the serving size (wt) were made based on the serving size patterns considered in EFSA Panel on Biological Hazards et al. (2018a) to obtain a better value for each population group.

In Guilherme (2012), 64% of the consumers were between 15 and 59 years of age (**Table 1**), therefore the Guilherme (2012) minimum, mode and maximum serving size (wt) were attributed to female 15-24, 25-44 and 45-64 age groups. The EFSA Panel on Biological Hazards et al. (2018a) serving size average for female 15-24, 25-44 and 45-64 age groups was calculated. Then, the serving size for other female age groups was transformed to a percentage based on that average, on EFSA Panel on Biological Hazards et al. (2018a) data (equation 1). With these percentages, it is possible to calculate the serving size for all the female age groups on Guilherme (2012) data (equation 2 applied for minimum, mode and maximum).

$$Serving\ size_{age,female} = \frac{EFSA_{age,female} \times 100}{average(EFSA_{15-24,female}, EFSA_{25-44,female}, EFSA_{45-64,female})}\ \% \qquad 1$$

$$wt_{age,female} = \frac{Guilherme_{15-64,female} \times Serving\ size_{age,female}}{100}\ grams \qquad 2$$

The male serving size was then transformed to a percentage based on the female serving size on EFSA Panel on Biological Hazards et al. (2018a) data (equation 4). With these percentages the males age groups' serving sizes were calculated on Guilherme (2012) data (equation 4 applied for minimum, mode and maximum). Results are present in **Table 2**.

$$Serving\ size_{age,male} = \frac{EFSA_{age,male} \times 100}{EFSA_{age,female}}\ \% \qquad 3$$

$$wt_{age,male} = \frac{Guilherme_{age,female} \times Serving\ size_{age,male}}{100}\ grams \qquad 4$$

Table 2 – Serving size by age and sex (wt_{pop}) based on Serra da Estrela cheese serving size in Guilherme (2012) and cheese serving size patterns used in EFSA Panel on Biological Hazards et al. (2018a).

Age Group	Sex	Min (g)	Most likely (g)	Max (g)
1-4	Female	7.06	9.41	28.24
	Male	6.66	8.88	26.64
5-14	Female	9.16	12.21	36.63
	Male	14.36	19.15	57.46
15-24	Female	15	20	60
	Male	16.42	21.89	65.66
25-44	Female	15	20	60
	Male	13.92	18.56	55.68
45-64	Female	15	20	60
	Male	14.25	19	56.99
65-74	Female	10.89	14.52	43.56
	Male	13.37	17.83	53.49
75+	Female	11.99	15.98	47.95
	Male	13.81	18.41	55.23

4.1.1.2. *L. monocytogenes* concentration

Information regarding the concentration of *L. monocytogenes* in the food product was also needed. As Guilherme (2012) did not present information on *L. monocytogenes* concentration in Serra da Estrela cheese, values from Gombas et al. (2003), also used for soft and semi-soft cheese in EFSA Panel on Biological Hazards et al. (2018a), were used for this QMRA, as Serra da Estrela cheese is classified as a semi-soft cheese. Values obtained from Gombas et al. (2003) were shown as a fitted cumulative distribution function. Cheese was modeled using beta-general distributions with a minimum of −1.69 $\log_{10}$ CFU/g, a maximum of 7 $\log_{10}$ CFU/g, α of 0.194 and β of 3.177 (EFSA Panel on Biological Hazards et al. 2018a).

4.1.1.3. Total number of eating occasions per year

The last parameter needed is the total number of eating occasions per year (TEO). To obtain these values, the mode of Serra da Estrela cheese size (wt_{cheese}) (**Table 1**) was divided by the most likely serving size value for each population (wt_{pop}) (**Table 2**) obtaining the number of servings per cheese (equation 5).

$$servings\ per\ cheese_{pop} = \frac{mode\ (wt_{cheese})}{mode(wt_{pop})} \qquad 5$$

Guilherme (2012) mentions that Serra da Estrela cheese can be consumed in 3 different occasions per year, mostly on Christmas and Easter, being eaten, in average, by 3 people. Taking this into consideration, it was assumed that, in average, a person eats Serra da Estrela cheese in 2 different occasions per year and that the cheese is eaten by 3 people. Therefore, it was considered that one person eats in average 0.67 cheeses per year (number of eating occasions divided by the number of average people eating the cheese). However, for younger individuals this is unlikely. Thus, it was considered that individuals with less than 15 years eat 0.25 cheeses a year. The average yearly number of people in Portugal was also calculated based on data on Portuguese demographic statistics from Eurostat (2021) which collects data from EU Member States on European demographic statistics. Population information was obtained for the period of 2012-2020, as it includes the most recent data and the period of Guilherme (2012) study. The mean was applied for each age to obtain the average number of individuals per age during this period. Afterwards, the sum between multiple ages was obtained to get the average population between this period for 7 age groups based EFSA Panel on Biological Hazards et al. (2018a) subgroups and the results are present on **Table 3**.

Table 3 - Portuguese yearly average population between 2012 and 2020 by age group and sex Eurostat (2021)

Age Group	Sex	Yearly average population between 2012 and 2020
1-4	Female	176893
	Male	185123
5-14	Female	490889
	Male	506086
15-24	Female	1423694
	Male	1348047

25-44	Female	1514148
	Male	1363804
45-64	Female	608225
	Male	637317
65-74	Female	604911
	Male	492655
75+	Female	645543
	Male	398074

With all this data, the TEO was calculated for each population group using the number of eating occasions per cheese per individual, the number of cheeses eaten per year per individual, and the number of people in that age group (equation 6 and **Table 4**). The TEO values obtained for each population are shown in Figure 5.

$$TEO_{pop} = servings\ per\ cheese_{pop} \times\ cheeses\ per\ year_{pop} \times number\ portuguese_{pop} \qquad 6$$

Table 4 - Description of the parameters used for the TEO calculation.

Parameter	Meaning	Distribution	Value	Source
wt_{cheese}	Serra da Estrela cheese weight	**Table 1**	**Table 1**	Guilherme (2012)
wt_{pop}	Serving size by population	**Table 2**	**Table 2**	Guilherme (2012)
Servings per $cheese_{pop}$	Number of servings per Serra da Estrela cheese by population	Constant	Equation 5	Guilherme (2012)
Cheeses per $year_{pop}$	Number of Serra da Estrela cheeses eaten per year by population	Constant	0.67 for ages ≥ 15 years 0.25 for ages < 15 years	Based on Guilherme (2012)
Number $portuguese_{pop}$	Portuguese yearly average population between 2012-2020 by age and sex	Constant	**Table *3***	Eurostat (2021)

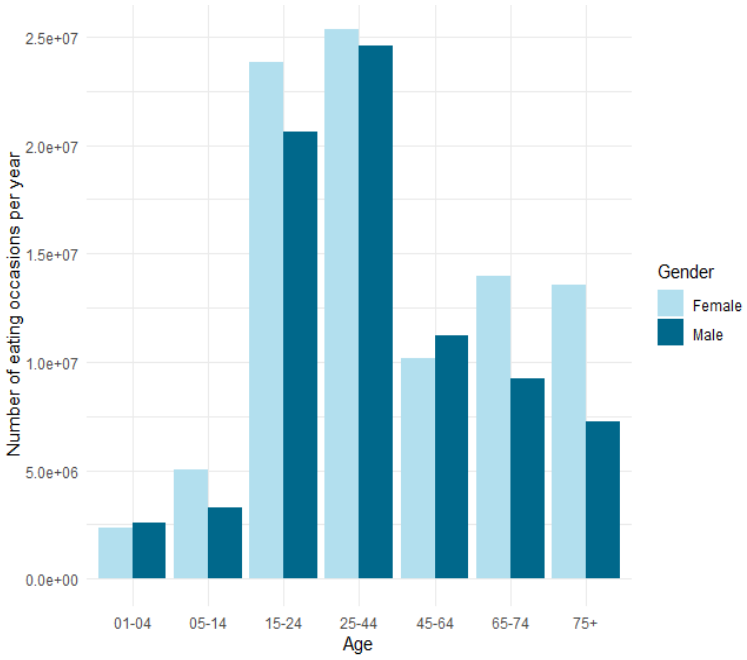

Figure 5 - Number of eating occasions per year (TEO) by age and gender.

4.1.2. WGS QMRA

In WGS QMRA, to obtain raw WGS data of *L. monocytogenes* strains in Serra da Estrela cheese, a search in multiple languages was performed on the National Center for Biotechnology Information (NCBI) Sequence Set Browser (NCBI 2021). Data was only found saved under the name “cheese” and “queso”. There were 755 samples available that were filtered to 514, by removing environmental and multi-ingredient samples, samples without a run number and duplicated samples. These samples came mostly from surveillance projects but also from research studies. Ideally, WGS data on *L. monocytogenes* from Serra da Estrela cheese would be used, however, since this data was not available, all the existing *L. monocytogenes* samples from cheese were used on the WGS QMRA to later make a conclusion for Serra da Estrela cheese based on similar cheeses. This way, a database was created based on WGS data from 514 *L. monocytogenes* strains isolated from cheese, including information on the project number, sample number, *L. monocytogenes* strain, cheese type, country and year of sample collection. To be able to download WGS data, the run number - a number that identifies the sample and allows download of the sample’s raw WGS data - for each sample was obtained from European Nucleotide Archive (ENA) Browser (ENA 2021) and added to the previous database.

Besides WGS data obtained from *L. monocytogenes* strains isolated from cheese, WGS data of *L. monocytogenes* with known clinical frequency was also needed. This data,

together with the respective run number was obtained from Maury et al. 2016. It consisted of 69 *L. monocytogenes* samples obtained from the French surveillance system in which the clinical frequency is given by the number of clinical isolates divided by the number of clinical isolates plus food isolates for a specific strain (Maury et al. 2016). A low clinical frequency means that the strain appears mostly in food and not in humans, indicating that it has low virulence whereas a high clinical frequency suggests the strain appears very frequently in humans, being considered highly virulent.

Using the run number WGS data without known clinical frequency and WGS data with known clinical frequency - a total of 583 samples - were downloaded to Computerome, a supercomputer for life sciences installed at Technical University of Denmark (DTU), using an in-house script.

Bioinformatic analysis and construction of machine learning models were performed using Danish National Supercomputer for Life Sciences, Computerome 2.0 (https://www.computerome.dk), a local server for a Linux-based command-line system. Computerome is accessible through the terminal MobaXterm. R version 4.0.0 (R Core Team 2020, Vienna, Austria) was used for statistical analyses. A scheme of this workflow is presented in Figure 6 and detailed in the following sections. This process allows to obtain all the needed data to build the predictive model and to make predictions.

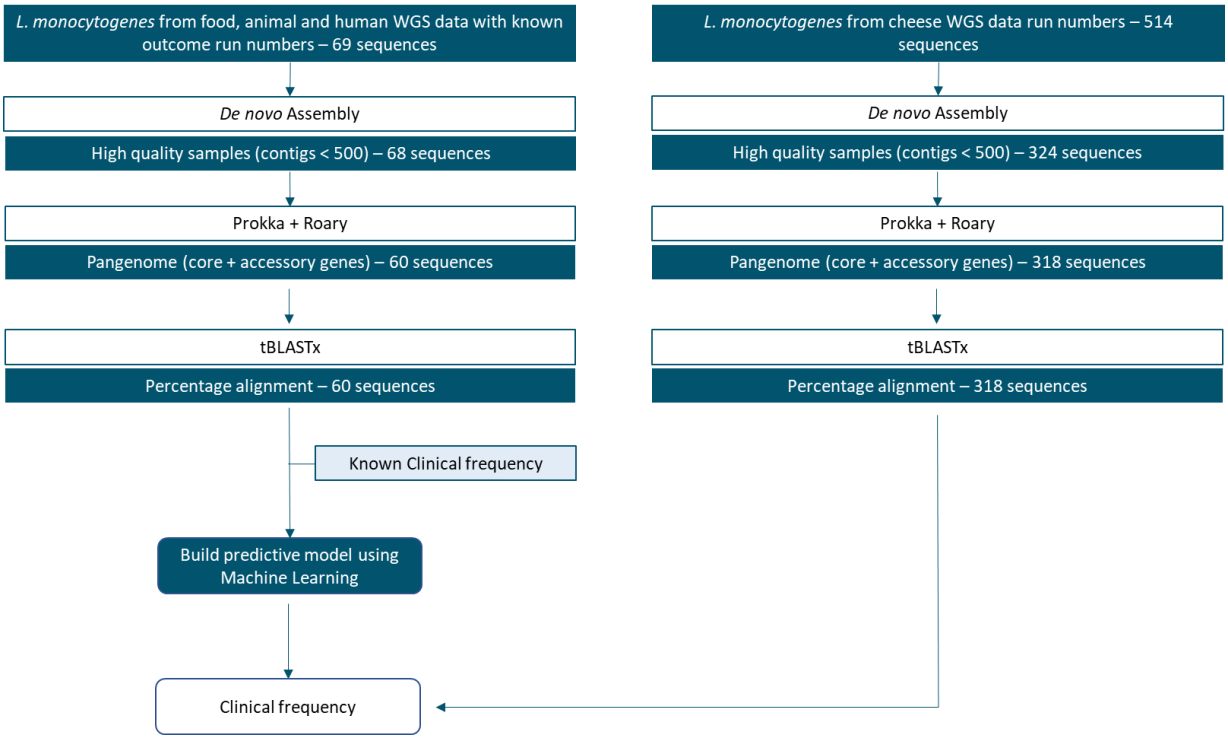

Figure 6 - Bioinformatics workflow and integration on predictive modelling.

4.1.2.1. Assembly

After downloading raw data, bioinformatic processes, explained in section 2.5., were applied, aiming for a DNA sequence analysis. The first step was the assembly to reconstruct the DNA sequence.

Raw reads, consisting of DNA fragments, were put together in continuous sequences, known as contigs. The process was done without the knowledge of a reference, also known as *de novo* assembled. The assembling was done using Food QC & Assembly pipeline that includes assembler SPAdes 3.9 (Bankevich et al. 2012). The quality of the assembly was assessed using number of contigs, N50, and total size of the assembly.

Genome assemblies with less than 500 contigs were kept in the dataset. N50 was verified for each sample to maintain only high N50 values. Eventually, the total size of the assembly was checked to match the expected size for a *L. monocytogenes* genome which is around 2.9 million bp.

Out of the 583 samples available for assembly, only 392 had high-quality (less than 500 contigs). This reveals the presence of sequences with extremely poor quality, published with alterations that hamper data usage.

4.1.2.2. Pangenome Analyses

The next step was to obtain the pangenome using all samples with high-quality assembly. Although results specify the presence (indicated with value 1) and the absence (indicated with value 0) of each gene, for the machine learning step, non-binary parameters are more accurate. Therefore, a good approach is to get the percentage alignment for each gene, instead of a 0/1 output. This is a percentage indicating the proportion of the gene present in the sample, which is equal to a reference of that gene. To obtain the percentage alignment needed for each gene, a basic local alignment search tool (BLAST) was used which found regions of similarities between the input sequence and the reference sequence.

The pangenomic analysis was performed on 392 assembled samples with high-quality. Annotation in some samples using Prokka failed, therefore results were only obtained for 378 samples.

After the pangenome analysis using Roary, the first output available was the summary statistics with the number of core and accessory genes, explained in section 2.5. **Figure 7** displays these results with each slice representing a gene category and correspondent number of genes. A total of 10,168 genes were found, 1,984 comprise the core genes including the soft-core (19.5%) and 8,184 the accessory genes (80.5%).

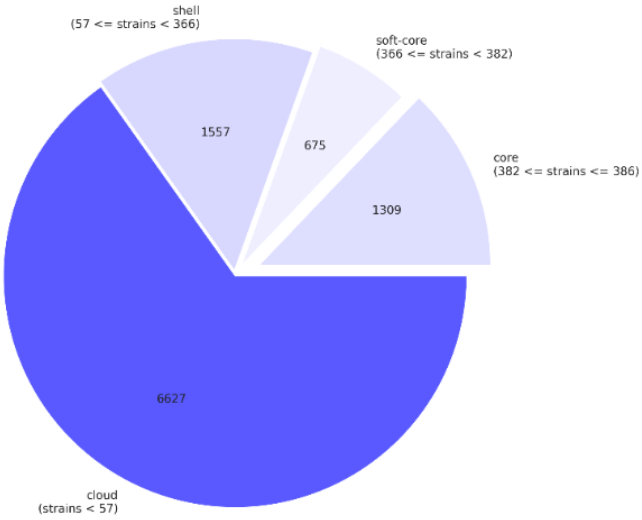

Figure 7 – Pangenome pie chart showing gene content obtained from Roary software.

The second output was the gene occurrence. In this output, samples corresponding genes are enumerated. As a *L. monocytogenes* reference sample with the identified genes was not used, Roary attributed a unique name to each gene. With this data a matrix was created (***Figure 8***) in which the dark blue represents the presence of a gene while, the light blue/white indicates the absence of a gene. In ***Figure 8***, it is also possible to visualize that, in the beginning (left side), each gene was present in every sample, meaning that those are the core genes, while in the right of the matrix, accessory genes are displayed.

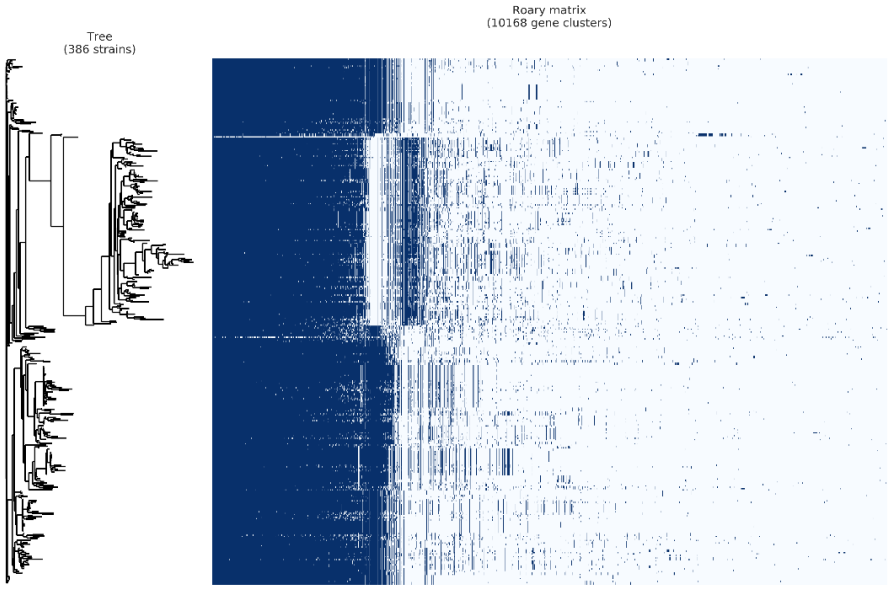

Figure 8 – Pangenome matrix from Roary that allows gene visualization.

Besides the above-mentioned outputs, multiple Rtab format files were obtained. To analyse these outputs, the code provided by Roary was used (Page et al. 2015) and results are displayed in **Figure 9** and **Figure 10**. In **Figure 9**, when the first sample (genome) is added, all the genes present in that sample are new. When a new sample is added, if the genes' repetitions are removed, the number of unique genes is obtained. As more samples are added, less new genes tend to appear, because all the samples have the core genes in common and only some accessory genes will appear as new genes. However, the number of unique genes tends to increase because of those accessory genes. In **Figure 10**, as expected, as new samples (genomes) are added, the number of total genes increases and the number of conserved genes, meaning genes present in all samples (core genes), tends to decrease, and then remain the same (variations considered artifacts).

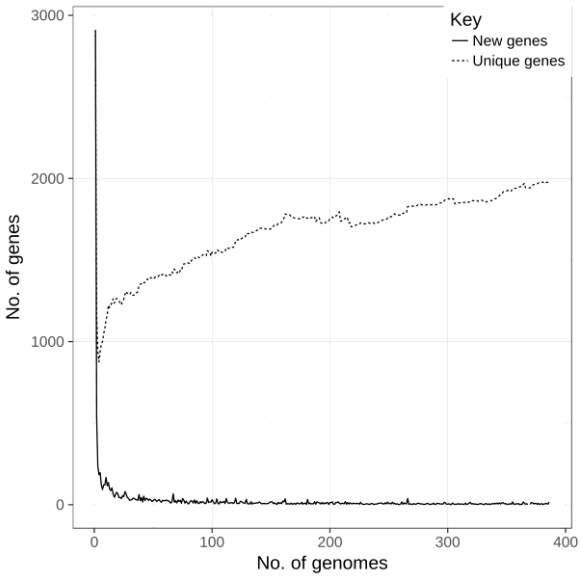

Figure 9 - Plot obtained from Roary outputs representing the number of new genes and the number of unique genes

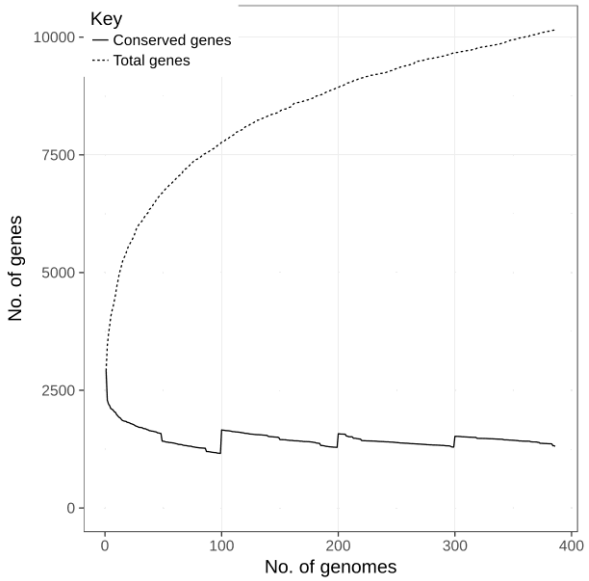

Figure 10 - Core pan plot obtained from Roary outputs representing the number of total genes and conserved genes.

Eventually, there were 60 samples with known clinical frequency available to develop a machine learning predictive model and 318 cheese samples, detailed in Appendix 2, to make predictions regarding their clinical frequency, using the predictive model. Each of the samples was described by 10,168 genes with the percentage alignment value.

4.2. Data Analyses

This section considers classic QMRA, in which the gQMRA model used to predict the probability of listeriosis in Portuguese population due to *L. monocytogenes* in Serra da Estrela cheese is described; and WGS QMRA, in which the construction of the machine learning model and its use to make predictions on clinical frequency of *L. monocytogenes* due to Serra da Estrela cheese is detailed.

4.2.1. Classic QMRA

With all the needed data for classic QMRA, *L. monocytogenes* gQMRA mentioned in section 2.4. was applied, R version 4.0.0 (R Core Team 2020, Vienna, Austria) was used to conduct all the analyses. As the purpose of this comparative study was to use one RTE food product as a case study, some adaptations of the gQMRA model had to be applied and are further detailed in the following sections. The gQMRA workflow is summarised on **Figure 11**.

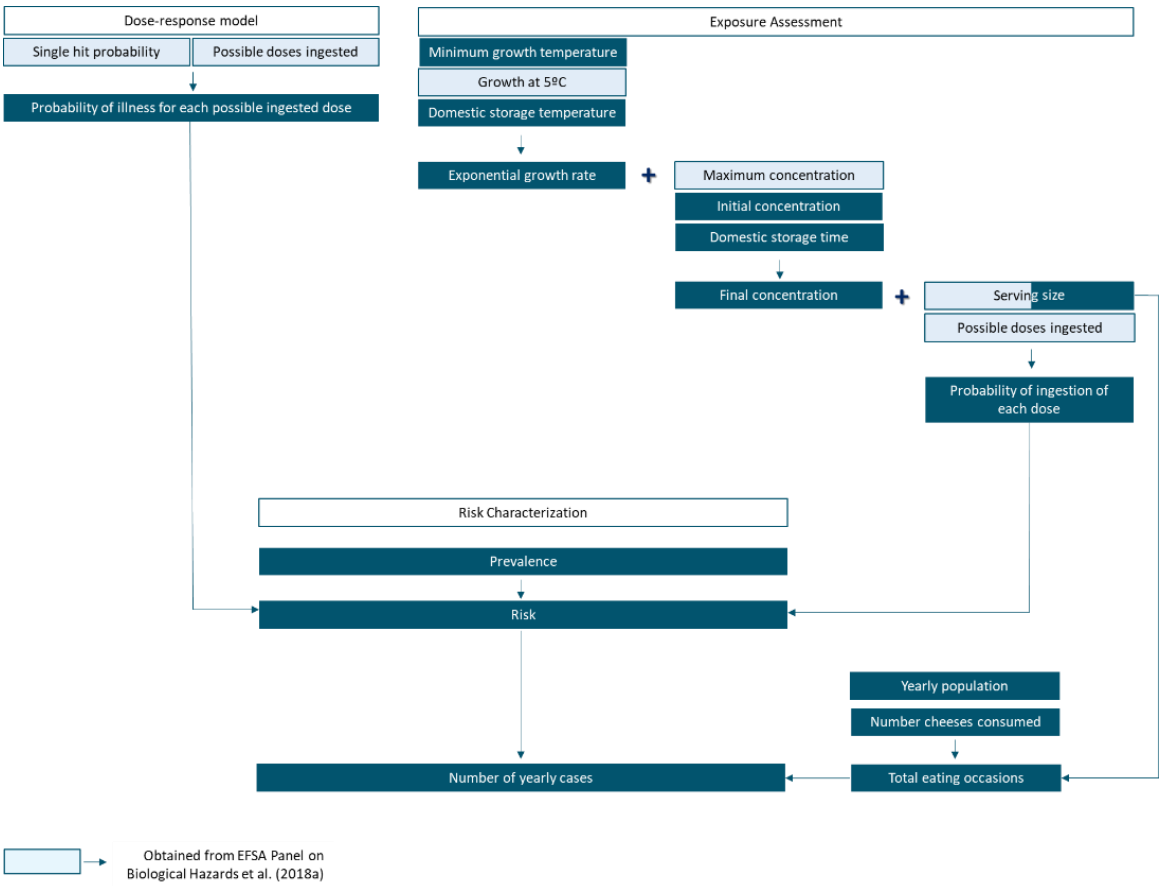

Figure 11 - gQMRA workflow.

4.2.1.1. Hazard identification

This step was already performed in section 2.3. in which *L. monocytogenes* importance is described, with special focus on RTE food products including cheese as used in this case study. However, it is important to add that because available data is related to retail and consumption and since the consumer phase is one of the most important ones as mentioned previously, the QMRA started at retail, focusing on the consumers' phase.

4.2.1.2. Hazard characterization (dose-response model)

The gQMRA developed by EFSA working group on *L. monocytogenes* contamination of RTE foods, uses the lognormal exponential model, based on Pouillot et al. (2015) (equation 7), for the dose-response model in the hazard characterization step (EFSA Panel on Biological Hazards et al. 2018a). This dose-response model is a single hit model as each cell is capable of initiating illness with low probability. Therefore, the model assumes no threshold, meaning that no matter how low the dose, there is always, at least in a mathematical sense, a non-zero probability of infection and illness. It also assumes an independent action, which means that the probability of a pathogen to cause an infection is independent of the number of pathogens inoculated as there is no interaction between cells. The basis of this model is the exponential dose-response model where the probability of illness for a specific dose is calculated as: $P_{ill}(\lambda) = 1 - e^{-r\lambda}$. From this basis, Pouillot et al. (2015) included the probability density of the r variable (f(r)), described as a lognormal distribution. Therefore, the model used in gQMRA gives the mean probability of illness for each dose by population. The variables present in the model are explained in **Table 5**.

$$P_{ill,pop}(\lambda) = 1 - \int_0^1 e^{-r\lambda} f(r) dr \qquad 7$$

Table 5 - Description of the parameters used for the dose-response model.

Parameter	Meaning	Distribution	Value	Source
r	Single hit probability which is the probability of one *L. monocytogenes* cell to successfully initiate illness	Lognormal	**Table *6***	Pouillot et al. (2015) and EFSA Panel on Biological Hazards et al. (2018a)

λ	Dose of *L. monocytogenes* ingested	-	0 to 12 $\log_{10}$ CFU by increments of 0.1	EFSA Panel on Biological Hazards et al. (2018a)

Regarding the "r" parameter, two parameters describe the lognormal distribution: the mean and the standard deviation. For the "r" mean calculation, the exposure model output, TEO per population and the average of the annual observed cases of listeriosis per population in the EU between 2008 and 2011 were used in EFSA Panel on Biological Hazards et al. (2018a). Estimating "r" mean consists of solving equation 8 where the only unknown value is mean of "r" as the standard deviation of "r" is considered constant and equal to 1.62 for all the populations (EFSA Panel on Biological Hazards et al. 2018a). In this equation $g(\lambda)$ describes the variability of the expected dose and $f(r)$ describes the variability of the single hit probability. In this QMRA, EFSA Panel on Biological Hazards et al. (2018a) values were used (**Table 6**).

$$Cases = TEO \times \left(1 - \int_{\lambda=0}^{\infty} \int_{r=0}^{1} e^{-r\lambda} g(\lambda) f(r) d\lambda dr\right) \quad 8$$

Table 6 - "R" parameter values for each age group by sex (EFSA Panel on Biological Hazards et al. 2018a).

Age Group	Sex	R mean	R standard deviation
0-4	Female	-14.5737	1.62
	Male	-14.4668	
5-14	Female	-14.916	
	Male	-15.0046	
15-24	Female	-14.3249	
	Male	-15.0357	
25-44	Female	-14.0246	
	Male	-14.7638	
45-64	Female	-14.0808	
	Male	-14.0446	
65-74	Female	-13.702	
	Male	-13.5598	
75+	Female	-13.5362	
	Male	-13.5358	

Regarding the "λ" parameter, as in EFSA Panel on Biological Hazards et al. (2018a), the possible ingested doses were defined from 0 to 12 $\log_{10}$ CFU by increments of 0.1. Therefore, 121 possible ingested doses of *L. monocytogenes* were analysed. With all the data available, the dose-response model was applied obtaining the probability of illness for each of the 121 possible ingested doses.

4.2.1.3. Exposure assessment

The first step in the exposure assessment is the calculation of the exponential growth rate (EGR) of *L. monocytogenes* from retail to consumption. Equation 9 is used in gQMRA model (EFSA Panel on Biological Hazards et al. 2018a) and its variables are detailed in **Table 7**.

$$EGR(T) = EGR(5ºC) \times \left(\frac{T - T_{min}}{5 - T_{min}}\right)^2 \qquad 9$$

$$if\ T < T_{min} \rightarrow ERG(T) = 0$$

Table 7 - Description of the parameters used to calculate the EGR.

Parameter	Meaning	Distribution	Value	Source
T_{min}	Minimum temperature for *L. monocytogenes* to grow	Constant	-1.18ºC	FDA and FSIS (2003) and EFSA Panel on Biological Hazards et al. (2018a)
EGR(5ºC)	Growth of *L. monocytogenes* at 5ºC	Lognormal	Mean = 0.0103 SD = 0.0151 Max = 0.0296 Min = 0	Pérez-Rodríguez et al. (2017) and EFSA Panel on Biological Hazards et al. (2018a)
T	Domestic storage temperature of Serra da Estrela cheese	Pert	Min = -0.5ºC Mode = Mean = 17ºC (assumption that the mode is equal to the mean) Max = 34.5ºC	Guilherme (2012)

Min = minimum; Max = maximum; SD = standard deviation

The model obtained random values for EGR at 5ºC from the truncated distribution. The same was performed for domestic storage temperature using a PERT distribution. With all data defined, EGR(T) equation was applied with the condition that if domestic storage temperature is below the minimum growth temperature, then EGR will be zero.

After knowing the growth rate, final concentration (C(t)) in CFU/g at the end of storage time was calculated using Rosso equation (equation 10) (EFSA Panel on Biological Hazards et al. 2018a). The equation variables are described in **Table 8**.

$$C(t) = \frac{C_{max}}{\left(1 + \frac{C_{max}}{C_{(0)} - 1}\right) \times e^{-ERG(T) \times t}} \qquad 10$$

Table 8 - Description of parameters used to calculate *L. monocytogenes* final concentration per gram of cheese.

Parameter	Meaning	Distribution	Value	Source
C_{max}	Maximum concentration or maximum population density of *L. monocytogenes*	Constant	7.28 $\log_{10}$ CFU/g	Pérez-Rodríguez et al. (2017) and EFSA Panel on Biological Hazards et al. (2018a)
$C_{(0)}$	Initial concentration of *L. monocytogenes* before storage	Beta-general	Min = -1.69 $\log_{10}$ CFU/g Max = 7 $\log_{10}$ CFU/g Shape 1 = 0.194 Shape 2 = 3.177	Gombas et al. (2003) and EFSA Panel on Biological Hazards et al. (2018a)
EGR(T)	Exponential growth rate at storage temperature T	Equation 9	-	From the previous step
t	Domestic storage time of Serra da Estrela cheese	Pert	Min = 1 day Mode = 9 days Max = 21 days	Guilherme (2012)

The original gQMRA model used data on the shelf-life after purchase and the proportion of remaining shelf life spent until consumption in order to obtain the domestic storage time. However, as Guilherme (2012) study provided information regarding domestic storage time, random samples were obtained from a PERT distribution of data. Random samples were also obtained from the beta distribution of the initial concentration data. With all the available data, the Rosso equation was applied and *L. monocytogenes* concentration at the time of consumption was obtained.

Finally, to obtain the probability of ingesting each dose of *L. monocytogenes* for each population ($P_{ing,pop}$), gQMRA used equation 11 in which the doses ingested by serving are calculated, equation 12 where a cumulative distribution function (F_x) is applied and equation 13 where the probability of ingesting a specific dose in one serving is calculated based on the cumulative distribution function results (EFSA Panel on Biological Hazards et al. 2018a). Equation variables are described in **Table 9**.

$$\lambda = C(t) \times wt_{pop} \quad 11$$

$$P_{ing,pop}(\lambda \leq x) = F_{\lambda}(x) \quad 12$$

$$P_{ing,pop}(x) = P_{ing,pop}(\lambda \leq x) - P_{ing,pop}(\lambda \leq x - 0.1) \quad 13$$

Table 9 - Description of the parameters used to calculate the probability of ingesting each dose.

Parameter	Meaning	Distribution	Value	Source
C(t)	Concentration of *L.monocyogenes* at consumption time	Equation 10	-	From the previous step
wt_{pop}	Serving size in grams	Pert	**Table 2**	Guilherme (2012) and EFSA Panel on Biological Hazards et al. (2018a)
x	A value for the dose parameter (λ)	-	A value between 0 and 12 $\log_{10}$ CFU by increments of 0.1	EFSA Panel on Biological Hazards et al. (2018a)

As data regarding serving size (wt_{pop}) variation was available, a new step was added to the original gQMRA in order to obtain random values for the serving size from the PERT distribution. With all available data, the probability of ingesting each of the 121 possible doses was obtained.

4.2.1.4. Risk characterization

The risk is the probability of illness per serving. It is calculated as the probability of ingesting a certain dose times the probability of illness from that dose. This value is then multiplied by the prevalence of *L. monocytogenes* in Serra da Estrela cheese (equation *14* and **Table 10**). The number of expected listeriosis cases per year for each population is obtained by the multiplication of the risk per eating occasion in that population with the TEO for that population (equation 15 and **Table 10**) (EFSA Panel on Biological Hazards et al. 2018a).

$$Risk_{pop} = \sum_{\lambda=0}^{120} [P_{ing,pop}\left(\frac{\lambda}{10}\right) \times P_{ill,pop}\left(\frac{\lambda}{10}\right)] \times P \qquad 14$$

$$Prediction\ of\ Yearly\ Cases = Risk \times TEO \qquad 15$$

Table 10 - Description of the parameters used to calculate cases per year per age and sex.

Parameter	Meaning	Distribution	Value	Source
$P_{ing,pop}(\lambda)$	Probability of ingesting dose (λ) for that population	Equation 11, 12 and 13	-	From the previous step
$P_{ill,pop}(\lambda)$	Probability of illness due to ingestion of dose (λ) for that population	Equation 7	-	From previous steps
P	Prevalence of *L. monocytogenes* in Serra da Estrela cheese	Constant	39.6%	Guilherme (2012)
TEO	Total number of eating occasions per year for that population	Constant	**Figure *5***	Guilherme (2012)

4.2.1. WGS QMRA

After obtaining the database associating *L. monocytogenes* strains to clinical frequency and after collecting WGS data from *L. monocytogenes* strains isolated from cheeses, the next step was to find patterns on the database to later predict the clinical frequencies on the cheeses. This process is called predictive modelling, which is a mathematical process that aims to predict future outcomes based on existing patterns in available data. In this study, a database was used, where the clinical frequency of the strains was known, to train the machine learning algorithm. The algorithm detected patterns between the percentage alignment for each of the 10,168 genes (that were the features for the machine learning model) and the clinical frequency (which was the outcome for the machine learning model). By knowing these patterns, when a new strain was inserted in the model, it evaluated the percentage alignment present in each gene of the strain and based on them, it predicted the clinical frequency. After building the model, ideally, WGS data obtained from *L. monocytogenes* isolated from Serra da Estrela cheese would be inserted in the model to make a prediction on clinical frequency. However, this data was not available, so clinical frequencies were predicted in multiple cheeses with different characteristics. By analysing the clinical frequencies predicted for cheeses similar to Serra da Estrela cheese, a conclusion might be taken for Serra da Estrela. A scheme of the workflow is represented in **Figure 12** and further detailed in the next sections. The model and predictions were made using R version 4.0.0 (R Core Team 2020, Vienna, Austria) and caret package.

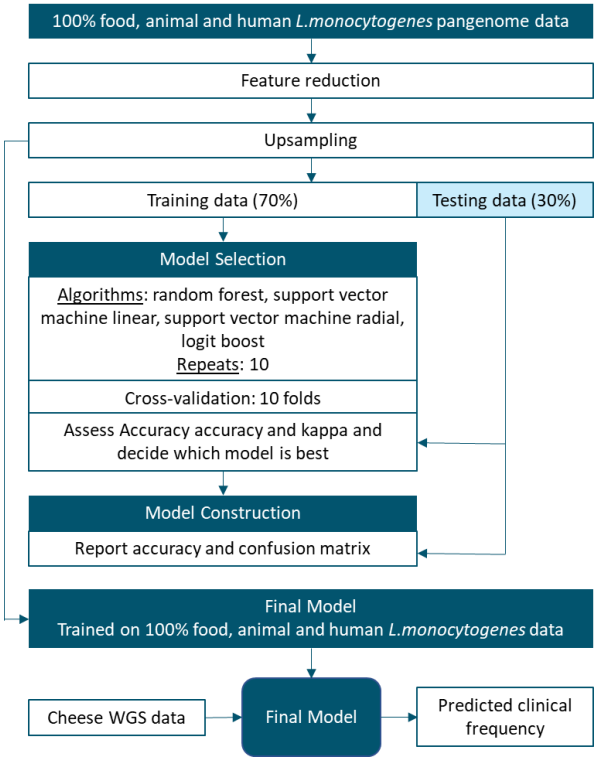

Figure 12 - Predictive modelling workflow.

4.2.1.1. Pre-processing

To link the WGS data with the clinical frequency, machine learning was used where the features are the percentage alignment for each gene and the outcome was the clinical frequency. Supervised learning was used as there was a labelled dataset and the goal was for the algorithm to make predictions based on that set of examples. Even though the outcome variable (clinical frequency) was continuous, a classification method was chosen, as the dataset available had gaps in some ranges of the outcome variable, hampering the development of a good model using regression. Therefore, the first step was data transformation, in order for it to be suitable for the classification method. The outcome variable, which is the clinical frequency, was transformed into four equal classes (**Figure 13**). The features, which are the percentage alignment for each gene, had an original range from 0 to 100 that was changed for 0 to 5.

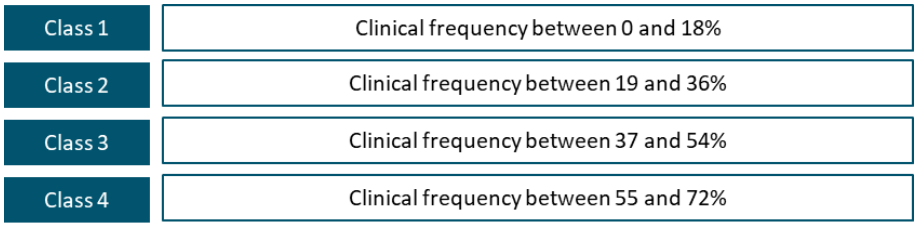

Figure 13 - Clinical frequency classes.

After data preparation, feature reduction was performed in order to eliminate zero and near-zero variance features, including absent features and features that were always present in samples or that did not have significant variance being irrelevant for the predictions.

4.2.1.2. Subsampling

Since the number of samples for each clinical frequency class was variable, a subsampling step was needed. Subsampling is a method that aims to avoid the negative impact that class imbalance has on model fitting. Subsampling examples are downsampling, in which random sampling will be made on the classes with more samples, so that all the classes have the same number of samples, as the class with the lowest number of samples. Another example is upsampling where classes with the lowest number of samples will be randomly sampled with replacement, so that all classes have the same number of samples as the class with the highest number of samples (**Figure 14**). Hybrid methods can also be

considered, in which downsampling is done in classes with more samples and upsampling is performed in classes with less data (Kuhn 2019). Through experiments on data, upsampling was chosen to balance classes.

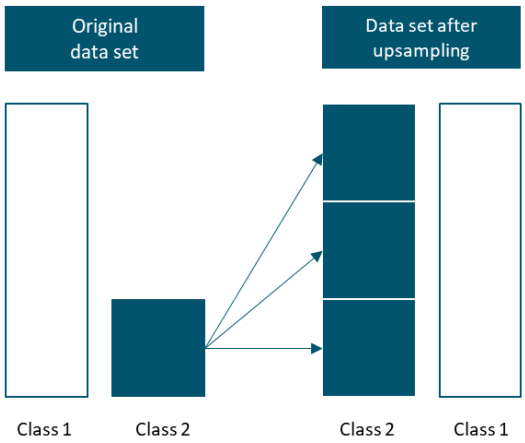

Figure 14 – Upsampling method used to balance dataset classes.

4.2.1.1. Data splitting and model training

After subsampling, the data was split into training set - corresponding to 70% of data with known clinical frequency, and testing set - corresponding to the other 30%. The training set was used to train four common and widely used machine learning methods (RF, SVM linear, SVM radial and LB) using a cross-validation of 10 folds. This means that the training data was divided into 10 equally sized parts called folds and the model was trained, using one of the folds called training data and evaluated using the other fold which is the testing data. After training the model with all the 10 possible splits, the final model was obtained, and it was tested with the testing set to obtain the accuracy and kappa values (**Figure 15**). In order to obtain more reliable results, this process was performed 10 times for each model, obtaining 10 accuracy and 10 kappa values for each of the 4 models.

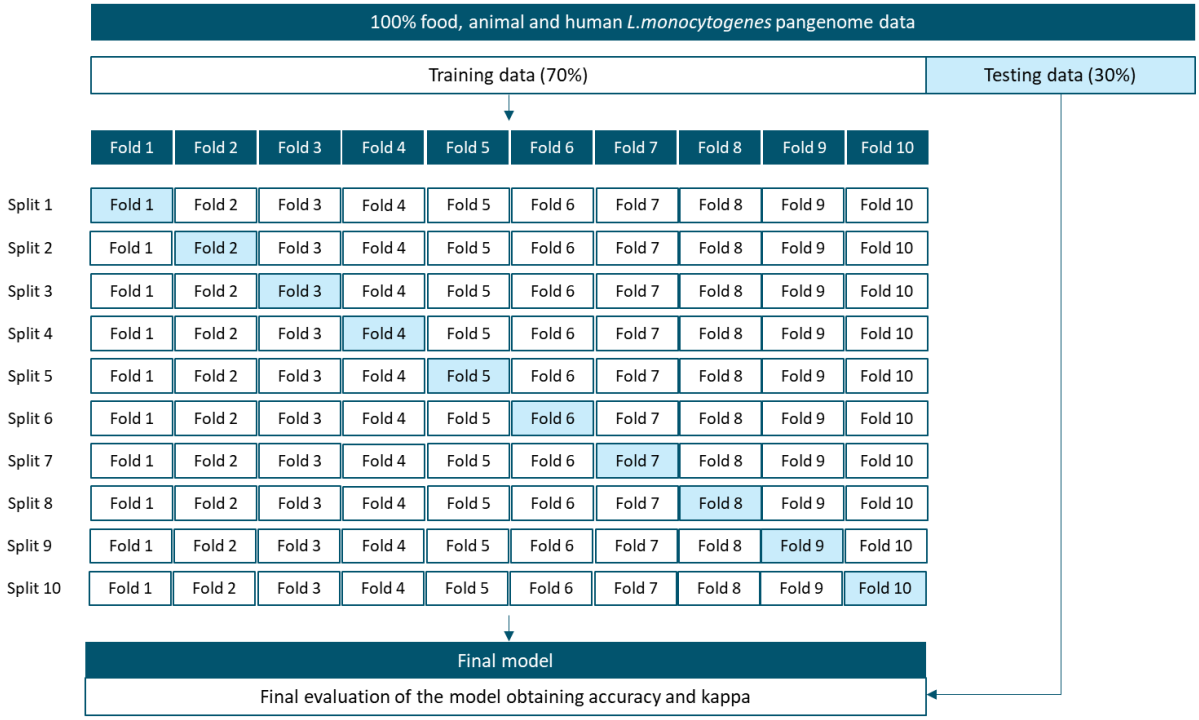

Figure 15 - Data splitting and model training.

4.2.1.2. Model selection, evaluation and final model

Accuracy and kappa values for each run were obtained, as well as the average for each parameter. After checking that the data met the assumptions behind the statistical method, an ANOVA test was conducted to see if differences between models were significant. Based on these results, with special focus on accuracy and kappa values, the best model was selected and evaluated. After being evaluated, this best model was trained as the final model to be used for predicting the clinical frequency of each *L. monocytogenes* obtained from cheese, by including all the upsampled data with known outcome and not just the training set (**Figure 16**).

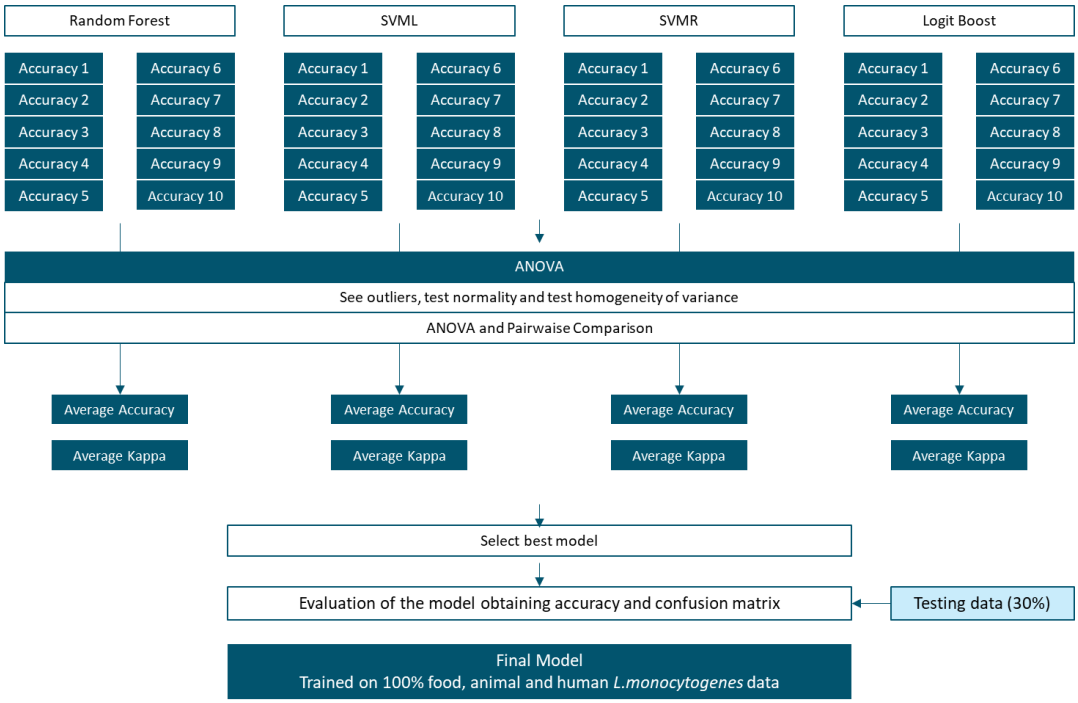

Figure 16 - Model selection and evaluation.

4.2.1.3. Variable importance and predictions

After training the final model, variable importance was investigated to identify genes that are more important in predicting the clinical frequency. However, since a reference *L. monocytogenes* was not used in the bioinformatic analysis, conclusions in this matter would require additional bioinformatic analysis that were not made. The final model was applied to data with unknown clinical frequency to make predictions regarding clinical frequency. By applying the model, a clinical frequency class was attributed to each cheese sample (**Figure 17**). Then, cheeses with similar characteristics to Serra da Estrela cheese – semi-soft, cured, made with raw milk, with sheep origin, from Portugal - were identified and a clinical frequency was attributed to *L. monocytogenes* present in Serra da Estrela cheese based on the clinical frequencies of the *L. monocytogenes* in those similar cheeses.

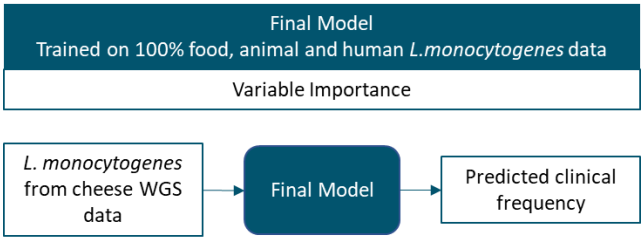

Figure 17 – Application of the final machine learning model for variable importance and clinical frequency predictions.

5. Results

In this section, the probability of listeriosis and the expected number of listeriosis cases in Portuguese population groups due to *L. monocytgenes* on Serra da Estrela cheese obtained in classic QMRA and the expected clinical frequency of *L. monocytgenes* due to its presence in Serra da Estrela cheese obtained in WGS QMRA are described.

5.1. Classic QMRA

5.1.1. Dose-response model

In the second step of classic QMRA where a dose-response model was developed to define the probability of illness from ingesting each dose by population groups, results obtained are illustrated in **Figure 18**. As expected, as the dose of ingested *L. monocytogenes* increases, so does the probability of illness. The probabilities of illness are similar between sexes for each age group. This is due to the fact that the single hit probability mean (r mean) is similar for each sex in the same age group. The biggest differences are observed at the 15-24 and 25-44 age groups where the probability of illness for females is higher than that for males. It is in these age groups that a bigger difference between the single hit probability mean (r mean) is present, being higher for females. Results also reveal that the probability of illness is below 1% when *L. monocytgenes* countings are between 9 to 10 $\log_{10}$CFU for all populations. An increasing trend in the probability of illness can be observed if *L. monocytgenes* countings are higher than the above-mentioned values. As an example, the increase in probability of illness found for 5 to 6 $\log_{10}$CFU of ingested *L. monocytgenes* is smaller than the one found for 10 to 11 $\log_{10}$CFU, which is expected as the dose-response model is exponential.

In both sexes the probability of illness is higher for the 75+ age group. Also due to the single hit probability (r mean), there is a tendency for the probability of illness to increase with age although there are some exceptions. The most significant one is at the 01-04 age group where the probability of illness is higher than for the 05-14 age group in females and higher than for the 05-14, 15-24 and 25-44 age groups in males.

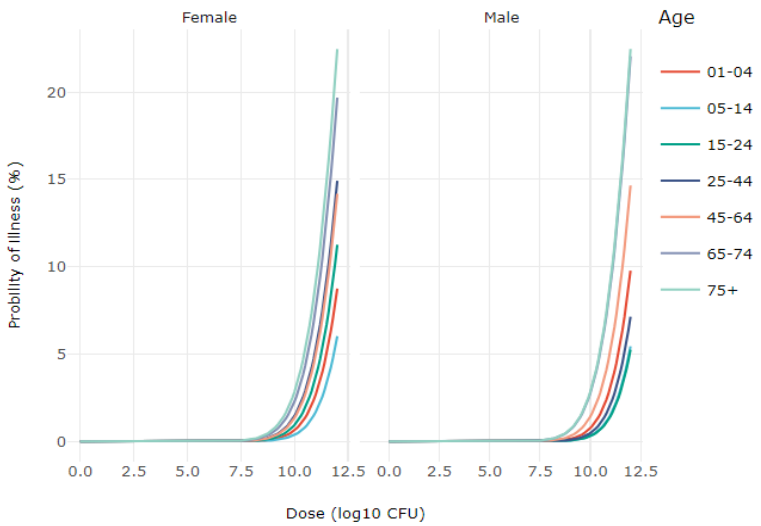

Figure 18 - Probability of illness after ingestion of each dose of *L. monocytogenes* by age and gender.

5.1.2. Exposure assessment

In the QMRA exposure assessment step, the probability of ingesting each dose was calculated for each population (**Figure 19**). Results reveal that the probability of ingesting each dose of *L. monocytogenes* is very similar for all population groups, as *L. monocytogenes* concentration is the same for each population, and the variation of serving size (wt_{pop}) does not seem to be sufficient to lead to big differences between populations nor sexes. The probability of ingestion of 0 log_{10}CFU of *L. monocytgenes* is around 50% and decreases immediately reaching the 6% at 0.1 log_{10}CFU. From around 8.8 log_{10}CFU the probability of ingestion becomes very close to 0% for all age groups and both sexes.

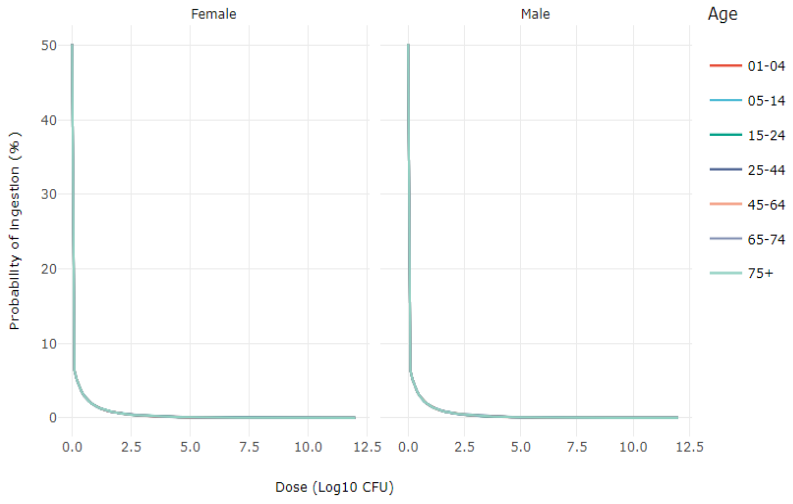

Figure 19 - Cumulative probability of ingesting each dose of *L. monocytogenes* by age and gender

5.1.3. Risk Characterization

Finally, in the risk characterization step the risk of illness was calculated (Figure 20). The risk of illness tends to be higher for females below the age of 45, and then the risk tends to be higher for males. In females the risk lowers from 01-04 to 05-14 age groups, rising for the 25-44 age group, reducing again for the 45-64 age group, increasing again, and reaching the highest level at 75+. For males, the same initial decrease was found but from 05-14 age group until the 75+, the risk of illness increased steadily, achieving its maximum at 75+.

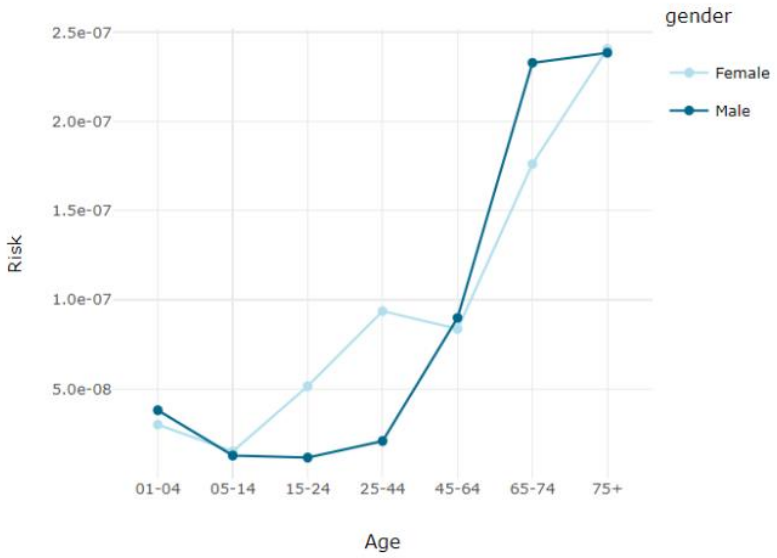

Figure 20 - Risk of illness by age and gender.

Classic QMRA final results are presented in **Table 11** and ***Figure 21***. According to this model, a total of 16 listeriosis cases per year are expected in Portugal due to the consumption of Serra da Estrela cheese. The predictions indicate that female sex is expected to be more affected with a total of 10 cases representing more than half of the expected cases. The most affected population is expected to be 75+ females with 3.27 cases, followed by 65-74 females, 25-44 females, 65-74 males and 75+ males. The model predicted 0.29 cases for children under 15 years.

Table 11 - Results from the classic QMRA on Serra da Estrela cheese using the gQMRA model.

Prevalence	Population Age	Population Sex	TEO	Risk per serving	Cases per year
0.3956	1 to 4	Female	9396125	3.12×10^{-08}	0.07
		Male	10422611	3.80×10^{-08}	0.10
	5 to 14	Female	20102287	1.50×10^{-08}	0.08
		Male	13211963	1.26×10^{-08}	0.04
	15 to 24	Female	71184700	5.07×10^{-08}	1.21
		Male	61590027	1.18×10^{-08}	0.24
	25 to 44	Female	75707400	9.27×10^{-08}	2.35
		Male	73476565	2.08×10^{-08}	0.51
	45 to 64	Female	30411250	8.24×10^{-08}	0.84
		Male	33548012	8.89×10^{-08}	1.00
	65 to 74	Female	41665267	1.75×10^{-07}	2.44
		Male	27633102	2.28×10^{-07}	2.11
	75 plus	Female	40387828	2.41×10^{-07}	3.27
		Male	21624153	2.38×10^{-07}	1.73

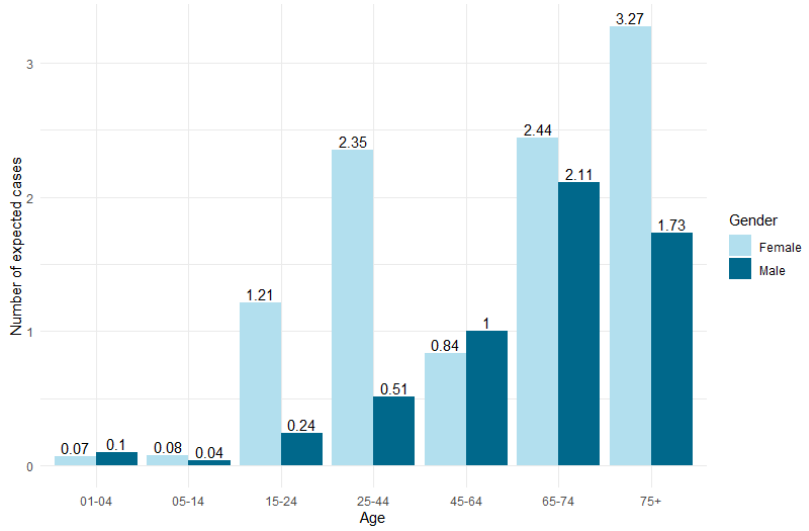

Figure 21 - Number of expected listeriosis cases in Portugal due to *L. monocytogenes* in Serra da Estrela cheese.

5.1.4. Uncertainties

As in any QMRA model, there are potential sources of uncertainty in some variables and methods applied. Uncertainty is present due to data and knowledge gaps. These potential sources are explained in **Table 12**. An important source of uncertainty is the dose response model as it was based on EFSA Panel on Biological Hazards et al. (2018a) gQMRA model

results and data. Data gaps are mostly present in the initial concentration (C_0) of *L. monocytogenes* in Serra da Estrela cheese and in cheese consumption.

Table 12 - Potential sources of uncertainty in the gQMRA and its potential impact on the final come based on EFSA Panel on Biological Hazards et al. (2018a).

Component of assessment	Assumption/data used	Source of uncertainty	Effect on expected cases
Prevalence	Obtained from a sample of 91 Serra da Estrela cheese from Guilherme (2012) study.	Performance of detection methods	Higher or Lower
		Competition with background flora	
		Representativeness of Serra da Estrela cheese	
		Outdated data	
Initial Concentration (C_0)	Obtained from EFSA Panel on Biological Hazards et al. (2018a) where Gombas et al. (2003) data from the United States of America on cheese was used, assuming a beta-general distribution with a minimum equal to −1.69 and maximum equal to 6.1.	Performance of detection and quantification methods	Higher or Lower
		Competition with background flora	
		Representativeness of Serra da Estrela cheese	
Domestic storage time	Obtained from a sample of 39 family aggregates from Guilherme (2012) study.	Representativeness of consumers	Higher or Lower
		Outdated data	
Domestic storage temperature	Obtained from a sample of 39 family aggregates from Guilherme (2012). Guilherme (2012) data was obtained during summer while the biggest consumption is in Christmas and Easter. The mode was considered to be the same as the mean and a PERT distribution was used. A constant refrigeration temperature was assumed (but variable between consumers). Storage time and temperature were considered independent factors.	Mode may not be equal to the mean	Higher or Lower
		Representativeness of consumers	
		Representativeness of biggest consumption period	
		Differences between temperatures on the surface and core of the cheese	
		PERT distribution may not be appropriate	
		Outdated data	
		Temperature conditions are dynamic and not constant	

		Storage time and temperature are expected to be dependent factors	
Growth	Minimum growth temperature was assumed to be -1.18ºC. It was assumed that the EGR is log-normally distributed based on EFSA Panel on Biological Hazards et al. (2018a). Data from Pérez-Rodríguez et al. (2017) was used for the probability distribution. Lag time was considered to be over at the retail level and no interaction with background flora was assumed A constant value for the maximum concentration was assumed	T_{min} value	Higher or Lower
		Lognormal distribution for EGR may not represent all sources of variability	
		Values of EGR distribution parameters	
		Lag time may not be completed from production to retail	
		Background flora may affect growth	
		The maximum concentration can vary depending on temperature, initial concentration, and background flora	
Consumption	An adaptation of the consumption values from Guilherme (2012) was made based on the EFSA Panel on Biological Hazards et al. (2018a) values to obtain appropriate values by population group. TEO was calculated assuming a 500g average Serra da Estrela cheese size, the serving size mode for each population group, the assumption that each individual ingests 1 or 2 cheeses a year according to their age and the yearly average population in Portugal	The adaptation made on the consumption values may not be the most accurate	Higher or Lower
		Representativeness of consumers	
		Serra da Estrela cheese size	
		Number of cheeses ingested yearly by individual	
		Variability in serving size and total number of eating occasions per year	
Dose response	Mean “r” values from the EFSA Panel on Biological Hazards et al. (2018a) study based on the output of the EFSA’s exposure model, average of the annual observed cases of listeriosis per subpopulation between 2008 and 2011 and TEO per subpopulation The “r” standard deviation was assumed to be the same for each subpopulation.	The “r” values used may not be the most accurate for the conditions of this study	Higher or Lower

Taking into account some of the uncertainties present in QMRA and in order to test mitigation strategies, alternative scenarios were tested evaluating the influence of some important variables in the number of expected cases (**Table 13**).

In order to understand the influence of domestic storage time, four alternative scenarios where tested. If the cheese is consumed in a maximum of 15 days maintaining the 9 days mode, the number of expected cases per year drops from 16 to 13. If a maximum and mode of 7 days is considered, the number of expected listeriosis cases drops to 9.4 cases. If the maximum is kept at 7, but the mode is 5, drops to 8.1 could be expected, and with mode of 3 the number drops to 7.1. As expected, this shows that the younger the cheese at the time of consumption, the lower the risk and the number of listeriosis cases. However, it is important to notice that competition with background flora and changes in the characteristics of the cheese according to the conservation temperature are not taken into account.

Regarding domestic storage temperature, a 2020 study investigated household fridge temperatures in multiple countries, considering three different age categories. As an alternative scenario to Guilherme (2012) domestic storage temperature data, the three fridge temperature hypotheses found in Portuguese households in Dumitrașcu et al. (2020), were tested. Therefore, gQMRA was run considering that Serra da Estrela cheese was always kept in the fridge, with i) a minimum temperature of 4.1ºC, maximum of 5ºC and mean of 5ºC; ii) a minimum temperature of 3.2ºC, maximum of 8ºC and mean of 5ºC; iii) a minimum temperature of 3.8ºC, maximum of 9.1ºC and mean of 6.4ºC (Dumitrașcu et al. 2020). For all these three hypotheses, the number of expected listeriosis cases per year would drop from 16 to a value close to 6. These results show that storing Serra da Estrela cheese in the fridge is a good mitigation strategy and the different fridge temperatures verified in Portuguese households do not significantly change this mitigation.

Finally, regarding consumption, the serving size was calculated attributing EFSA Panel on Biological Hazards et al. (2018a) serving size values to females as explained on section 2.1.1. In this part of the study the gQMRA model was run attributing EFSA Panel on Biological Hazards et al. (2018a) serving size values to the males. This also resulted in 16 expected cases per year with very small changes in the number of expected cases for each population group. In another analysis, by changing the number of cheeses eaten per year per individual to 0.5 for people younger than 15 years and to 1 for people with 15 years or more, the number of expected cases per year goes up from 16 to 24 cases and by changing them to half, the number of expected cases is reduced to 7.5 listeriosis cases per year. Since Serra da Estrela cheese size varies from 0.5kg to 1.7kg and the 0.5kg were used to obtain the TEO, three other possibilities were tested in this part of the study to see how this variable influences the expected number of cases per year. By changing the cheese size to 700g the number of

expected cases per year rises to 23, increasing it to 1kg, the number of cases rises to 32 and increasing it to 1.7kg the number of expected listeriosis cases is 55.

Table 13 - Alternative scenario testing to evaluate the changes in the expected number of yearly listeriosis cases.

Parameter	Value	New Value	New number of yearly cases	Observations
Prevalence (P)	39.6%	0.42%	0.17	As expected, if the prevalence is lower, so is the risk and the number of expected cases
Domestic storage time (t)	Min = 1 day Most likely = 9 days Max = 21 days	Min = 1 day Most likely = 9 days Max = 15 days	13	As expected, the older the cheese consumed, the higher the number of expected cases
		Min = 1 day Most likely = 7 days Max = 7 days	9.4	
		Min = 1 day Most likely = 5 days Max = 7 days	8.1	
		Min = 1 day Most likely = 3 days Max = 7 days	7.1	
Domestic storage temperature (T)	Min = -0.5ºC Mean = 17ºC Max = 34.5ºC	Min = 4.1ºC Mean = 5ºC Max = 5ºC	5.9	Refrigerated temperatures are a good mitigation strategy
		Min = 3.2ºC Mean = 5ºC Max = 8ºC	5.9	
		Min = 3.8ºC Mean = 6.4ºC Max = 9.1ºC	6.2	
Serving size (wt_{pop})	Table **2** which is based on females	Based on males	16	Almost the same expected number of cases per population

Number of cheeses eaten per year	0.25 below 15 years old 0.67 above 15 years old	0.5 below 15 years old 1 above 15 years old	24	As expected, the more cheese eaten, the higher the expected cases
		0.125 below 15 years old 0.31 above 15 years old	7.5	
Cheese weight (wt_{cheese})	0.5kg	0.7 kg	23	
		1 kg	32	
		1.7 kg	55	

5.1. WGS QMRA

5.1.1.1. Upsampling

Due to the existence of imbalanced classes observed in **Figure 22**, upsampling was performed obtaining balanced classes as seen in **Figure 23**. After this process all possible clinical frequency classes have 19 samples, allowing the development of a more accurate model to make predictions.

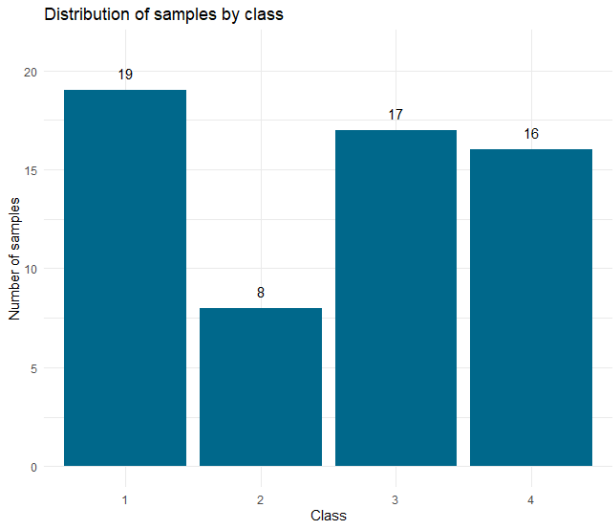

Figure 22 - Distribution of samples by class before upsampling.

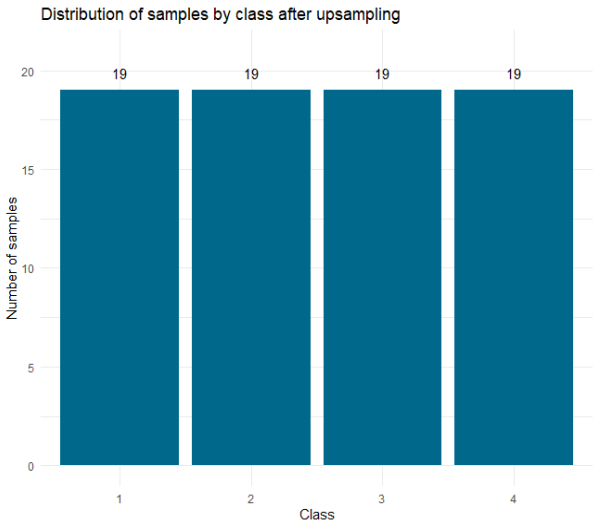

Figure 23 - Distribution of samples by class after upsampling.

5.1.1.2. Model selection

Boxplots built to visualise the minimum, maximum, median, interquartile range and outliers for the 10 accuracy and kappa values obtained for each model are present in **Figure 24** and **Figure 25**. In both parameters, LB is the model with better values, with an average accuracy of 86.3% and an average kappa of 81.6%, followed by SVM linear with an average accuracy of 84% and an average kappa of 78.7%, RF with average accuracy of 83% and an average kappa of 77.3%, and SVM radial with average accuracy of 75% and an average kappa of 66.7%.

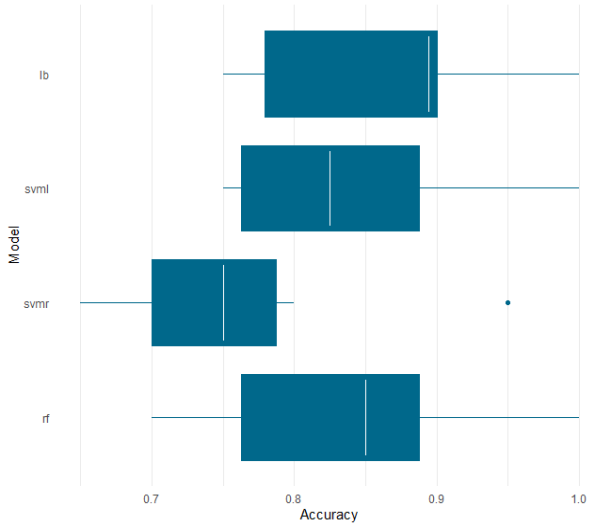

Figure 24 - Accuracy per model.

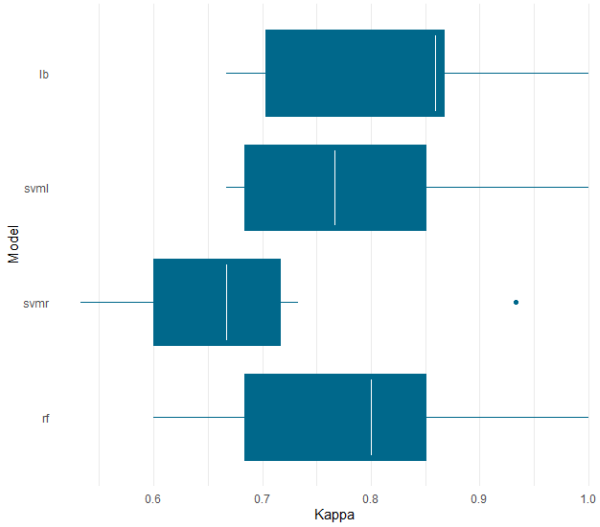

Figure 25 - Kappa per model.

After assuring that there were no extreme outliers and that data met the assumption of normality and homogeneity of variance, an ANOVA test was performed to confirm if the differences between accuracies for each model were significant. Because a p-value of 0.043, was obtained, the null hypothesis was rejected, and at least one of the differences in accuracy between models was significant. Therefore, a pairwise comparison test was performed, and results indicated that a significant difference between LB and SVM radial models ($p<0.05$) (**Figure 26**).

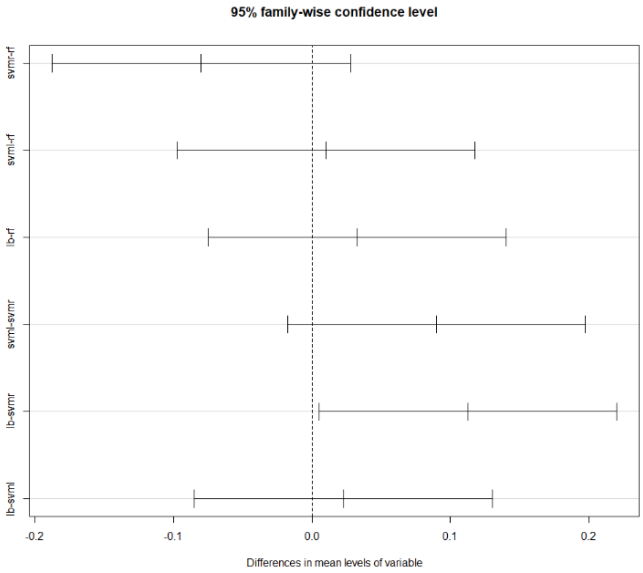

Figure 26 - Pairwise comparison test results (LB – Logit Boost, RF – Random Forest, SVML – Support Vector Machine Linear, SVMR – Support Vector Machine Radial).

Based on these results, LB model was selected to be used as the final model for clinical frequency predictions on *L. monocytogenes* isolates with unknown clinical frequency.

5.1.1.3. Model evaluation

To evaluate LB model a confusion matrix was built (**Figure 27**). Sensitivity and specificity presented values higher than 0.85, with the exception of class 1 sensitivity. The balance accuracies for each class were all above 0.8 (**Table 14**). Accuracy was 90% (95 % CI: 68.3%, 98.77%) and kappa was 86.67%. The no information rate (NIR), which is the accuracy achievable by always predicting the majority class label, was 0.25, and the p-value for the test to checked if accuracy was higher than NIR was 1.611×10^{-9}, indicating the null hypothesis rejection and concluding that accuracy is higher than NIR.

Table 14 - Sensitivity, specificity and balanced accuracy for Logit Boost.

Class	1	2	3	4
Sensitivity	1	1	0.6	1
Specificity	1	1	1	0.867
Balanced Accuracy	1	1	0.8	0.933

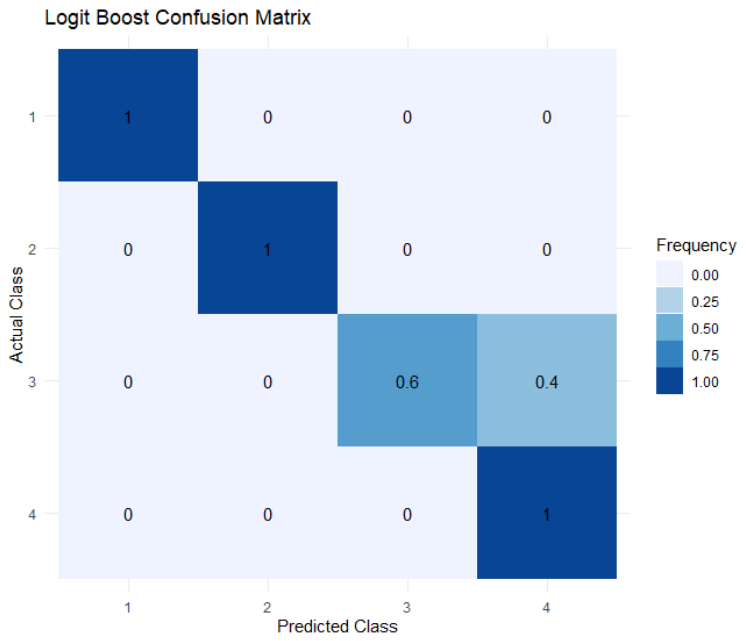

Figure 27 - Logit Boost confusion matrix.

5.1.1.1. Variable importance

The variable (genes) importance was analysed, and the 20 most important variables are described in **Figure 28**. Genes 6518, 4254 and 3413 have probabilities above 0.8 as important predictors for all classes. Using a reference *L.monocytgenes* pangenome could be beneficial so that the function of these genes, if known, may enable further discussions of their roles.

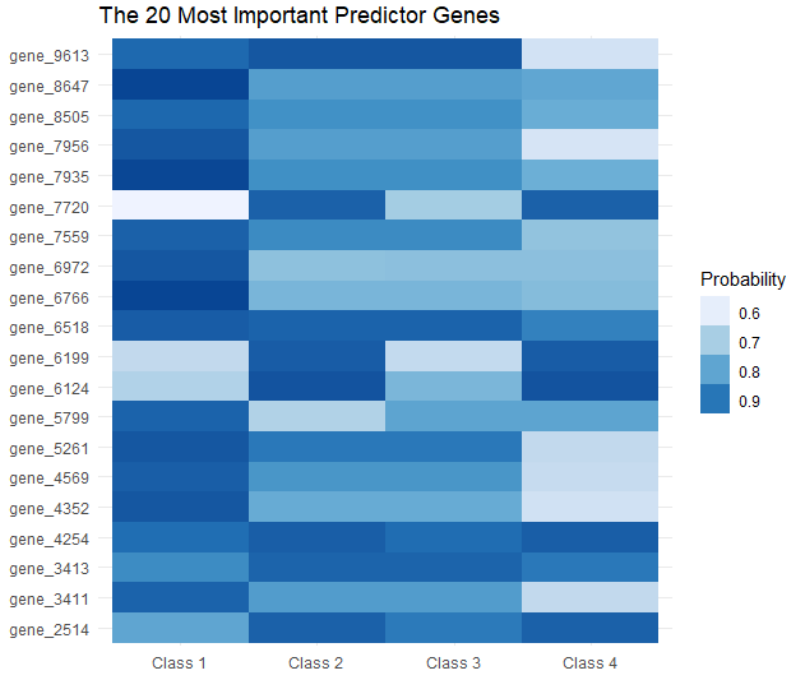

Figure 28 - Twenty most important predictor genes.

5.1.1.2. Prediction

Based on the LB model trained with all the data with known clinical frequency, predictions were made for all the samples with unknown clinical frequency. **Figure 29** illustrates the distribution of the predictions by clinical frequency class. A summary of the cheeses' types for each class is made on **Table 15**. Challenges were faced because, in some cases, the source of the *L. monocytogenes* WGS data did a poor description of the cheese type from which the *L. monocytgenes* was isolated from. This association between cheeses and clinical frequency allows a prior assumption of the clinical frequency, and therefore virulence, when a *L. monocytogenes* strain is identified in a specific cheese. However, some cheeses can be associated with multiple clinical frequencies as they might be contaminated with different *L. monocytogenes* strains.

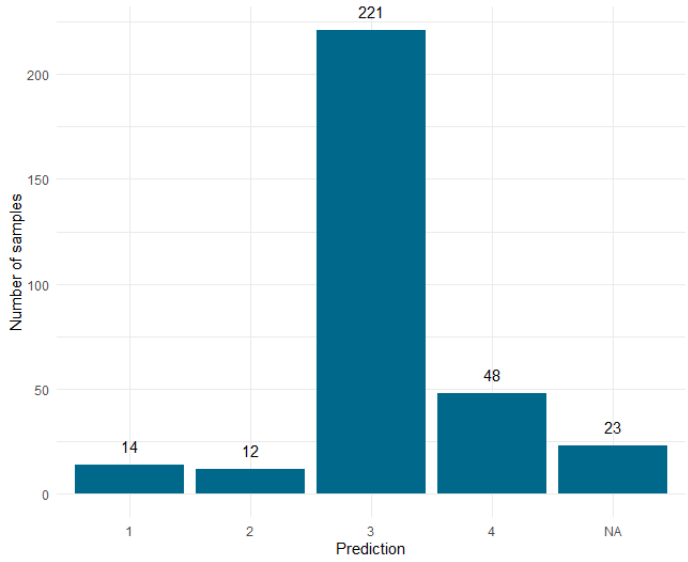

Figure 29 - Distribution of the samples by clinical frequency class (1 - 0 to 18%, 2 - 19 to 36%, 3 - 37 to 54%, 4 - 55 to 72%, NA - no prediction).

Table 15 - Cheeses' characteristics per clinical frequency class.

Clinical Frequency	Cheese Type	Country	Oldest Year	Most Recent Year
0-18	Cream Mold-ripened Blue-veined Fresh curd Robiola Pineta Brie-style Gouda wheel Blue crumbles Cubed Cheddar Le vigneron marc Gorgonzola Others: raw milk	Chile France Italy USA	2001	2020

19-36	Shredded Cheddar Curd White Mexican Fermier goat Camembert goat Others: raw milk, cow	Italy Israel USA	2002	2020
37-54	Ricotta Salvadorian String Fresh Curd Spanish Style Cheddar balls Fontina Mexican Soft White Aged raw milk Cotija Soft ripened American curd Bocconcini Cream Robiola Pineta Fresh Morbier aged 60 days Oaxaca string Moliterno al tartufo Gorgonzola dolce cheese Shredded Mozarella Cubed cheddar Monterey Jack Blue Swiss Asadero Burrata soft Brie Bucheron goat Talleggio Fermier goat Mozarella Others: raw milk, raw milk aged 60 days, soft, hard, semi soft, hard aged, soft ripened, goat, cow/sheep	Australia Bulgaria Canada Chile France Greece Israel Italy Mexico Poland Portugal Spain Uruguay USA	1905	2020
55-72	Fontina Blue Mexican Fresh Asadero Chihuahua type Others: raw milk, soft, hard, goat, pasteurized cow milk	Chile France Italy Mexico USA	2001	2021

In **Table 15**, all cured cheeses are in the 37-54 clinical frequency category and the same goes for the cheeses from Portugal. Besides, a lot of semi-soft cheeses are in this category as well as cheeses from sheep. Raw milk cheeses and cheeses between 2003 and 2004 are present in all categories. Bearing in mind that Serra da Estrela cheese is a cured sheep semi-soft cheese obtained from raw milk, Serra da Estrela cheese is expected to have a clinical

frequency between 37 and 54%. As in classic QMRA, there are some uncertainties associated with data and methodology. These uncertainties are explained in **Table 16** and are mainly due to uncertainties associated with the samples with known clinical frequency.

Table 16 - Uncertainties of WGS QMRA data using machine learning.

Component of assessment	Assumption/data used	Source of uncertainty	Clinical frequency
Samples with known clinical frequency	Obtained from	Representativeness of strains	Higher or Lower
		Samples imbalances	
		Differences between population groups may exist for the clinical frequency	
Machine learning algorithm	LB, RF and SVM	Accuracy of other possible algorithms	Higher or Lower
Clinical Frequency	Machine learning outputs	Approximation for Serra da Estrela cheese	Higher or Lower

6. Discussion

Using classic QMRA, the yearly risk of listeriosis due to consumption of Serra da Estrela cheese was estimated, based on data from Guilherme (2012) and EFSA Panel on Biological Hazards et al. (2018) data and gQMRA model. The model predicted the risk of listeriosis per serving per population group and a total of 16 listeriosis cases in Portugal in one year due to the consumption of Serra da Estrela cheese with the elder groups being more affected. Multiple scenarios were tested showing mainly the importance of keeping the cheese in the fridge. Therefore, this method informs risk managers on the severity of the problem and gives clues for intervention measures.

In WGS QMRA based on available cheese genomic data, a machine learning model was trained with French *L. monocytogenes* samples with known clinical frequency. This model predicted a clinical frequency of 37 to 54% for Serra da Estrela cheese, also identifying the genes and cheeses associated the most with clinical cases. These outcomes may inform risk managers on the cheeses and *L. monocytogenes* strains for which interventions such as product withdrawal, are most needed. If a good surveillance system with sample collection is implemented, it can also promote the early detection and response to listeriosis outbreaks.

Both approaches allow to address different questions and may support different types of control measures. A summary of the comparison between approaches is present in **Table *17*** and further detailed in the following sections.

Table 17 - Summary of the characteristics of each QMRA approach.

	Classic QMRA	**WGS QMRA**
Data	Concentration of bacteria in food Prevalence of bacteria in food Serving size Consumption patterns Yearly population R mean and standard deviation Growth rate of bacteria at 5ºC Storage patterns	WGS data with clinical frequency known WGS data of the bacteria, ideally from the food source being studied
Steps	Problem formulation Hazard identification Data collection Dose-response model Exposure assessment Risk characterization Test alternative scenarios Uncertainties	Problem formulation Obtain WGS data Genome Assembly Pangenome Analysis Data preparation Feature reduction Subsampling Machine learning Predictions Prediction analysis Uncertainties
Skills	Expert knowledge about food product and methods R programming language for gQMRA	Bioinformatics Machine learning
Outputs	Dose-response model Risk per serving Expected number of yearly disease cases Test different scenarios and mitigation strategies	Most important genes for determining strain virulence Predict clinical frequency on new samples
Uncertainties	Concentration Prevalence Serving size Consumption patterns Growth of bacteria Storage patterns Dose-response	Samples with known clinical outcome Machine learning algorithm Clinical frequency attribution by cheese characteristics
Risk management main benefit	Good analysis to make decisions regarding mitigation strategies	Early detection of outbreaks Fast analysis to make decisions on strain specific interventions such as product withdrawal

6.1. Problem formulation

Problem formulation is the first step of any QMRA, to define the goal of the assessment.

In Appendix 1, it is revealed that the existing classic QMRA aimed to assess the expected listeriosis cases and to identify appropriate mitigation strategies in food products for which safety concerns exist, due to bad practices, frequent pathogenic bacteria isolation, intensive manipulation, and different microbiological criteria limits in food products. As in those studies, the goal in this study classic QMRA was to determine the expected number of listeriosis cases and mitigation measures that can be applied to reduce these numbers.

In the existing WGS QMRA, mentioned in Appendix 1, the goal was to evaluate the possible integration of WGS in classic QMRA to improve hazard identification, dose-response model or exposure assessment, and consequently risk characterization due to the fact that classic QMRA does not account for differences in virulence and growth between bacterial strains. Even though some studies mention the possible application of WGS QMRA independently of classic QMRA, such as Njage et al. 2018 and Njage et al. 2019, this study aimed to further explore the independent application of WGS QMRA. The goal was to assess the clinical frequency of *L. monocytgenes* and possible applications of these outputs.

Therefore, both QMRAs aimed to predict the frequency of listeriosis in humans due to *L. monocytogenes* in Serra da Estrela cheese.

6.2. Data Collection

One of the most important components to compare approaches is the required data to perform each type of QMRA. Data acquisition is an important step of QMRA since data quality influences the confidence on the results. When data is not available, the need for assumptions arises, creating uncertainties that complicate decision making for risk managers. Therefore, the aim is to acquire the most reliable and the most recent data so that risk managers can make the best decisions possible.

While developing classic QMRA, data on the considered food product and its consumption was needed. Data acquisition in this method is usually a time-consuming step, as literature studies, expert knowledge and available data must be used to acquire data needed for the analysis. This data is also difficult to obtain because it may involve access to confidential data, data with large variability associated and therefore dependent on a large number of samples to increase reliability and data that is difficult to be collected. This study was limited to the consumer phase, but in QMRAs where the production phase is also assessed, the amount of required data is even higher. This need for specific data leads to the

use of assumptions every time data is not available, or cannot be obtained, which consequently increases the uncertainties associated with the risk assessment.

On the other hand, the use of genomic data did not require data regarding the product and its consumption, which reduces uncertainties when compared with classic QMRA as the use of assumptions is less frequent. However, it required a database with WGS associated with clinical frequency to create a good model to make predictions. The creation of this database implied sampling from food and humans. This database was available from a previous study, however continuous update is important to keep the models as accurate as possible. Also, continuous surveillance and sampling are important to have real and updated clinical frequency values as it is influenced by clinical and the food isolates, *e.g.*, poor food sampling can lead to an overestimation of clinical frequency. Besides the database, WGS data of *L. monocytogenes* from samples similar to those from Serra da Estrela cheese were used, however, ideally, the samples should be from Serra da Estrela cheese itself to reduce uncertainties. It is also important that WGS data source is well described so that further conclusions regarding risk attribution can be made.

In the end, classic QMRA needs plenty of information and data, which makes this step the most time and cost consuming, and prompt to uncertainties. WGS QMRA needs less data and therefore less uncertainties are present. Although data is much more objective and less variable, a good systematic sample collection system associated with detailed metadata and NGS instruments are needed.

6.3. Data Preparation

After data collection, data preparation is needed in both approaches to obtain all the data in the appropriate format. Data preparation is a group of methods that will transform raw data in data that can be easily used for analysis.

After getting the required data for classic QMRA, data is usually ready to be used. However, in the case where assumptions from other data have to be considered, as the ones performed in this study for serving size, for example, some calculations are required. Once again, data preparation leads to uncertainties.

When using genomic data, data preparation step is vital. WGS data needs to be treated using bioinformatics as this process allows to obtain an input to be used to train the machine learning algorithms and to make predictions. However, different methods can be used to achieve usable data. In this work, the pangenome was used to have information regarding all genes present in *L. monocytgenes* strains. Other methods like cgMLST can also be used but involve less genes and consequently less features involved in the machine learning model. Independently of the chosen bioinformatic method, this bioinformatic phase requires multiple

steps that can take a long time to be completed, and above all requiring professionals with knowledge in the field. Therefore, in this method there is a shift of time and cost from data collection to data preparation and analysis, as mentioned in Pielaat et al. (2015). It is also important to mention that, when genomic data from previous studies is used, new obstacles emerge as it happened in this study. Multiple challenges were faced in the application of bioinformatic tools and methods to obtain valuable results from the available data. Even though data is published and with free access, its usage is made difficult by data modifications. Reproducibility of data and methods are advantages of using genomic data and machine learning, however it is crucial that data is shared in a user-friendly way. It would be important to have some data standardization, avoiding situations that increase the analysis time.

As in data collection, this step is prone to increase the uncertainties in classic QMRA method. When it comes to WGS QMRA data, this step is the most time and cost consuming phase, requiring the usage of multiple bioinformatic tools by specialists with the presence of obstacles when data is not shared in a standard way.

6.4. Data Analyses

With data collection and data preparation steps completed, the next phase is the data analysis where the goal is to go from data to results.

Classic QMRA has four well defined steps, and numerous equations based on different assumptions to reach results in each risk assessment step. Experience, literature, and expert knowledge is often needed to choose the best approach. This variety of approaches makes classic QMRA harder to be standardized and reproducible. However, gQMRA used in this study is a very helpful method because the model is already constructed and can be reproduced for multiple RTE food products. Standardized methods as this one can make QMRA faster. Besides, the usage of software such as R (R Core Team 2020, Vienna, Austria) used in the gQMRA model makes it easier to run the model multiple times and to take into account variability and uncertainty which can enhance reliability on results.

Regarding the analysis of genomic data, machine learning is used as a tool to achieve results. Therefore, besides bioinformatics, the team needs to include professionals with experience in machine learning. Beyond the human resources needed, it is also important to know that the algorithms take some time to be trained so that they can create models to be used to make predictions on new data. Once again, this shows the shift of cost and time from data collection to data analysis. However, once a good model is built, it is reproducible, the database can be constantly updated, and new samples can be inserted to make predictions.

In conclusion, the possible methods to be used in classic QMRA are vast and this type of risk assessment would benefit from standardized methods in software like R (R Core Team

2020, Vienna, Austria), whereas WGS QMRA would benefit from the fact that once the model is built, it can be updated and is constantly reusable. This step is a part of the most cost and time consuming in WGS QMRA, but time and cost can be reduced in future applications of the previously built model.

6.5. Results

With data analyses completed, plots, tables and other results were obtained that need to be correctly interpreted to achieve relevant conclusions to be used by risk managers.

In the performed classic QMRA, one of the first obtained results was the relation of ingested dose of bacteria and probability of illness. This is an output that was not obtained in the WGS QMRA, however, as mentioned in Appendix 1, studies have shown that WGS can be useful as a complement to this step in classic QMRA (Fritsch et al. 2018).

Another set of outputs, from the exposure assessment, involved predicting the probability of ingesting each possible dose of bacteria. Once again, this output was not obtained in WGS QMRA, but previous studies (Appendix 1) have shown that WGS data associated with phenotypic data regarding bacterial growth can contribute to a more precise exposure assessment (Fritsch et al. 2019; Njage et al. 2020).

Classic QMRA final output is the probability of illness per serving and the expected number of disease cases per year by population group. This output representing the number of people that will be affected by the bacteria is useful for risk managers to make decisions regarding the involved food product. Besides, the possibility of changing some parameters in classic QMRA to evaluate their influence on the risk and expected number of disease cases is extremely useful in risk management, so that uncertainties can be tested, and mitigation strategies can be defined.

When it comes to the WGS QMRA, it was possible to detect the genes that had more influence in clinical frequency, but because a reference *L. monocytogenes* was not used, it was not possible to ascertain if this study's results are in accordance with previous studies (Njage et al. 2018), or even to identify and characterize previously known or new genes of importance. Besides, the expected clinical frequency of *L. monocytgenes* due to multiple cheeses was obtained based on the genes of the *L. monocytogenes* strain taken from those cheese samples. This clinical frequency indicates the percentage of total isolates of that strain that were isolated from a human clinical case, which gives some information regarding strain's virulence. Therefore, with these results, when a specific cheese is contaminated or the risk of illness from that cheese is being evaluated, an expected clinical frequency can be predicted. Such prediction was made for *L. monocytgenes* in Serra da Estrela cheese, based on the data obtained in this study, that creates an association between clinical frequency and cheese type.

However, it is important to notice that some of the data used in the prediction had the cheese type detailed while other samples specified neither the cheese characteristics nor type, which hampered the association between the cheese type and the clinical frequency. Also, in cases where a new sample is available, it is possible to know precisely the expected clinical frequency based on the genes present on a particular *L.monoytogenes* strain in a sample, by inputting WGS sample data in the preciously built model.

Classic QMRA allows prediction of listeriosis cases by population group and alternative scenario testing giving important information to risk managers. Although results obtained in WGS QMRA are different and do not allow scenario testing, it is possible to predict the clinical frequency according to the contaminated cheese or according to WGS data of *L. monocytogenes*, allowing risk managers to take action as soon as possible.

6.6. Limitations and Future Perspectives

Previous studies (Appendix 1) have shown that WGS can be integrated in different steps of classic QMRA. This study clarified that, even though WGS data can be incorporated in classic QMRA to improve it, the two approaches performed separately address different questions and support different types of management decisions. Having a model build for predicting clinical frequency can be very interesting, as bacterial samples from food or humans as a result of surveillance networks, or from outbreaks, can be inserted in the model and its clinical frequency can be predicted. Based on the prediction, an outbreak can be detected and rapid measures, such as product withdrawal, can be applied if the predicted clinical frequency is high. If the predicted clinical frequency is not that high, it is possible to proceed to a classic QMRA to understand which mitigation measures should be applied to lower the risk.

It is important to mention that in this study *L. monocytogenes* and cheese were used as a case study, but gQMRA can be used for other RTE food products. The model built using machine learning can be used for other food products and machine learning models like this one can be built for other bacteria. Besides, other types of outcome variables can be used in the machine learning predictive model such as clinical frequency for each population group, or clinical outcome (*e.g.*, gastroenteritis, influenza like symptoms, abortion, or meningitis).

With the aim to outline the advantages, disadvantages and future usage of each method, a Strengths Weaknesses Opportunities and Threats (SWOT) analysis was made for each method and is presented in **Table 18** and **Table 19**.

Table 18 - SWOT analysis of classic QMRA.

Strengths	Opportunities
• Most developed method • Widely used • Gives the number of expected cases by population group • Allows alternative scenario and mitigation strategies testing	• Programming languages like R allow reproducibility as verified in the gQMRA model • WGS can be integrated in some steps
Weaknesses	**Threats**
• Need for a lot of information regarding the bacteria, the food product, the consumers, and the consumption patterns and the more specific the better • Based on assumptions when data is not available which leads to multiple uncertainties	• Genomic data comes as a more accurate and precise method with less assumptions

Table 19 - SWOT analysis of WGS QMRA.

Strengths	Opportunities
• Less use of assumptions becoming a more precise method • Once the model is built it can be continuously used to make new predictions • Allows early detection of outbreaks • Promotes withdrawals only when it is needed, avoiding unnecessary ones • Can be used to refine classic QMRA steps	• WGS is becoming cheaper (50-100€ now and maybe 10€ in the future (Cavaco and Leekitcharoenphon 2017)) • It allows reproducibility • Construction of a good surveillance system
Weaknesses	**Threats**
• Reliability on the results is dependent on the existence of a good database to train the model • For now, it is difficult to be used to test alternative scenarios and mitigation strategies	• Need for sample cultivation and the use of NGS to obtain WGS data • Need for data where the clinical frequency is known • Need for bioinformaticians Need for personal able to programme and use machine learning Time and cost consuming data analysis step • WGS data shared with alterations or incomplete is difficult to reuse

In future studies, it would be interesting to evaluate the efficiency of a WGS QMRA where a highly virulent strain was identified and an extreme measure like product withdrawal was taken, versus a classic QMRA where, for the same bacteria, only mitigation measures where applied, such as lower storage temperatures and shorter storage times. Assessing which method saves most health losses and which method costs the least, would make this

comparison of the approaches more detailed and future applications of each method more reliable. Besides, future research can also assess the possibility of building more WGS QMRA machine learning predictive models, more standardized models for classic QMRA, like gQMRA, for other bacteria and food products, and models for the integration between the two approaches.

7. Conclusion

This study disclosed that WGS QMRA and classic QMRA answer different risk assessment questions and support different types of risk management decisions. Their usage presents benefits for the risk management component of risk analysis. While using WGS QMRA enables to act quickly when needed, using classic QMRA allows to implement measures to mitigate the risk.

It was concluded that both approaches have space to grow in the future, together integrating WGS in classic QMRA, and separately with more standardized models for classic QMRA and more machine learning models for WGS QMRA.

Further studies on the risk management phase of risk analysis should be performed to define the good usage of WGS QMRA for decision making and the risk-benefit comparison between decisions based on WGS QMRA and decisions based on classic QMRA.

8. References

[ASAE] Autoridade de Segurança Alimentar e Económica. Listeria monocytogenes. [accessed 2021 May 4].https://www.asae.gov.pt/seguranca-alimentar/riscos-biologicos/listeria-monocytogenes.aspx.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 19(5):455–477. doi:10.1089/cmb.2012.0021.

Bassett J, Nauta M, Lindqvist R, Zwietering M. 2012. Tools for Microbiological risk assessment. ILSI Europe Report Series.

Bemrah N, Sanaa M, Cassin MH, Griffiths MW, Cerf O. 1998. Quantitative risk assessment of human listeriosis from consumption of soft cheese made from raw milk. Preventive Veterinary Medicine. 37(1–4):129–145. doi:10.1016/S0167-5877(98)00112-3.

Brownlee J. 2019. Machine Learning Evaluation Metrics in R. R Machine Learning. [accessed 2021 Jun 15]. https://machinelearningmastery.com/machine-learning-evaluation-metrics-in-r/.

Brusa V, Prieto M, Campos CA, Epszteyn S, Cuesta A, Renaud V, Schembri G, Vanzini M, Michanie S, Leotta G, et al. 2021. Quantitative risk assessment of listeriosis associated with fermented sausage and dry-cured pork shoulder consumption in Argentina. Food Control. 123:107705. doi:10.1016/j.foodcont.2020.107705.

Campagnollo FB, Gonzales-Barron U, Pilão Cadavez VA, Sant'Ana AS, Schaffner DW. 2018a. Quantitative risk assessment of Listeria monocytogenes in traditional Minas cheeses: The cases of artisanal semi-hard and fresh soft cheeses. Food Control. 92:370–379. doi:10.1016/j.foodcont.2018.05.019.

Cavaco L, Leekitcharoenphon P. 2017. Whole genome sequencing of bacterial genomes - tools and applications. Coursera. [accessed 2020 Dec 20]. https://www.coursera.org/learn/wgs-bacteria.

[CDC] Centers for Disease Control and Prevention. 2016. Whole Genome Sequencing (WGS). CDC. [accessed 2021 Apr 6]. https://www.cdc.gov/pulsenet/pathogens/wgs.html.

Chen J, Karanth S, Pradhan AK. 2020. Quantitative microbial risk assessment for Salmonella: Inclusion of whole genome sequencing and genomic epidemiological studies, and advances in the bioinformatics pipeline. Journal of Agriculture and Food Research. 2:100045. doi:10.1016/j.jafr.2020.100045.

Delua J. 2021. Supervised vs. Unsupervised Learning: What's the Difference? IBM. [accessed 2021 Jun 15]. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

Donges N. 2021. The Random Forest Algorithm: A Complete Guide. Built In. [accessed 2021 Jun 15]. https://builtin.com/data-science/random-forest-algorithm.

Duarte AS, Pamp S, Gompel L, Leekitcharoenphon P, Hald T, Munk P, Bortolaia V. 2018. Metagenomics applied to surveillance of pathogens and antimicrobial resistance. Coursera. [accessed 2020 Dec 5]. https://www.coursera.org/learn/metagenomics.

Dumitrașcu L, Nicolau AI, Neagu C, Didier P, Maître I, Nguyen-The C, Skuland SE, Møretrø T, Langsrud S, Truninger M, et al. 2020. Time-temperature profiles and Listeria monocytogenes presence in refrigerators from households with vulnerable consumers. Food Control. 111. doi:10.1016/j.foodcont.2019.107078.

[ECDC] European Centre for Disease Prevention and Control. Surveillance Atlas of Infectious Diseases. [accessed 2021 May 4]. https://atlas.ecdc.europa.eu/public/index.aspx.

EDES. 2012. Handbook 1.4 - Food Safety System: Risk Assessment. www.coleacp.org/edes.

[EFSA] European Food Safety Authority Panel of Biological Hazard, Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, Davies R, de Cesare A, Hilbert F, et al. 2019. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. EFSA Journal. 17(12):5898. doi:10.2903/j.efsa.2019.5898. [accessed 2021 May 7]. www.efsa.europa.eu/efsajournal.

[EFSA] European Food Safety Authority Panel on Biological Hazards, Ricci A, Allende A, Bolton D, Chemaly M, Davies R, Fernández Escámez PS, Girones R, Herman L, Koutsoumanis K, et al. 2018a. Listeria monocytogenes contamination of ready-to-eat foods and the risk for human health in the EU. EFSA Journal. 16(1):5134. doi:10.2903/j.efsa.2018.5134. [accessed 2021 May 3]. www.efsa.europa.eu/efsajournal.

[EFSA] European Food Safety Authority Panel on Biological Hazards, Ricci A, Allende A, Bolton D, Chemaly M, Davies R, Fernández Escámez PS, Girones R, Herman L, Koutsoumanis K, et al. 2018b. Jan 24. Listeria monocytogenes generic Quantitative Microbiological Risk Assessment (gQMRA) model. doi:10.5281/ZENODO.1117742. [accessed 2021 May 3]. https://zenodo.org/record/1117742.

Ekblom R, Wolf J. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications. 7:1026-1042. doi: 10.1111/eva.12178

[ENA] European Nucleotide Archive. 2021. ENA Browser. [accessed 2021 Apr 2]. https://www.ebi.ac.uk/ena/browser/text-search.

Eurostat. 2021. Data Explorer. [accessed 2021 May 4]. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_pjan&lang=en.

[FAO] Food and Agriculture Organization of the United Nations, [WHO] World Health Organization. 2003. Hazard Characterization for Pathogens in Food and Water: Guidelines. Rome Microbiological Risk Assessment Series Report No.: 3.

[FAO] Food and Agriculture Organization of the United Nations, [WHO] World Health Organization. 2008. Exposure assessment of microbiological hazards in food: Guidelines. Rome Microbial Risk Assessment Series Report No.: 7.

Fazil AM. 2005. A primer on risk assessment modelling: focus on seafood products. FAO Fisheries Technical Paper. 462:undefined. [accessed 2021 May 3]. http://www.fao.org/3/a0238e/a0238e00.htm.

Fritsch L, Felten A, Palma F, Mariet JF, Radomski N, Mistou MY, Augustin JC, Guillier L. 2019. Insights from genome-wide approaches to identify variants associated to phenotypes at pan-genome scale: Application to L. monocytogenes' ability to grow in cold conditions. International Journal of Food Microbiology. 291:181–188. doi:10.1016/j.ijfoodmicro.2018.11.028.

Fritsch L, Guillier L, Augustin JC. 2018. Next generation quantitative microbiological risk assessment: Refinement of the cold smoked salmon-related listeriosis risk model by integrating genomic data. Microbial Risk Analysis. 10:20–27. doi:10.1016/j.mran.2018.06.003.

Garrido V, García-Jalón I, Vitas AI, Sanaa M. 2010. Listeriosis risk assessment: Simulation modelling and "what if" scenarios applied to consumption of ready-to-eat products in a Spanish population. Food Control. 21(3):231–239. doi:10.1016/j.foodcont.2009.05.019.

Gombas DE, Chen Y, Clavero RS, Scott VN. 2003. Survey of Listeria monocytogenes in Ready-to-Eat Foods. http://meridian.allenpress.com/jfp/article-pdf/66/4/559/1675442/0362-028x-66_4_559.pdf.

Guilherme V. 2012. Contributo para uma avaliação de risco de Listeria monocytogenes em queijo Serra da Estrela. [Lisboa].

Huang L, Hwang CA. 2012. In-package pasteurization of ready-to-eat meat and poultry products. In: Advances in Meat, Poultry and Seafood Packaging. Elsevier Ltd. p. 437–450.

Kuhn M. 2019. Subsampling for Class Imbalances. The caret Package. [accessed 2021 Jun 16]. https://topepo.github.io/caret/index.html.

Mataragas M, Zwietering MH, Skandamis PN, Drosinos EH. 2010. Quantitative microbiological risk assessment as a tool to obtain useful information for risk managers - Specific application to Listeria monocytogenes and ready-to-eat meat products. International Journal of Food Microbiology. 141(SUPPL.):S170–S179. doi:10.1016/j.ijfoodmicro.2010.01.005.

Maury MM, Tsai YH, Charlier C, Touchon M, Chenal-Francisque V, Leclercq A, Criscuolo A, Gaultier C, Roussel S, Brisabois A, et al. 2016. Uncovering Listeria monocytogenes hypervirulence by harnessing its biodiversity. Nature Genetics. 48(3):308–313. doi:10.1038/ng.3501. [accessed 2021 Jun 11]. https://pubmed.ncbi.nlm.nih.gov/26829754/.

Nauta MJ. 2008. The Modular Process Risk Model (MPRM): a structured Approach to Food Chain Exposure Assessment. In: Schaffner D, editor. Microbial risk analysis of food. Washington DC: ASM Press. p. 99–136.

[NCBI] National Center for Biotechnology Information. 2021. Sequence Set Browser. [accessed 2021 May 10]. https://www.ncbi.nlm.nih.gov/Traces/wgs/.

Njage PMK, Henri C, Leekitcharoenphon P, Mistou MY, Hendriksen RS, Hald T. 2018. Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. Risk Analysis. 39(6):1397–1413. doi:10.1111/risa.13239.

Njage PMK, Leekitcharoenphon P, Hald T. 2019. Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic Escherichia coli. International Journal of Food Microbiology. 292:72–82. doi:10.1016/j.ijfoodmicro.2018.11.016.

Njage PMK, Leekitcharoenphon P, Hansen LT, Hendriksen RS, Faes C, Aerts M, Hald T. 2020. Quantitative microbial risk assessment based on whole genome sequencing data: Case of listeria monocytogenes. Microorganisms. 8(11):1–24. doi:10.3390/microorganisms8111772.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 31(22):3691–3693. doi:10.1093/bioinformatics/btv421.

Pedamkar P. 2020. Ensemble Techniques. EDUCBA. [accessed 2021 Jun 29]. https://www.educba.com/ensemble-techniques/.

Pérez-Rodríguez F, Carrasco E, Bover-Cid S, Jofré A, Valero A. 2017. Closing gaps for performing a risk assessment on Listeria monocytogenes in ready-to-eat (RTE) foods: activity 2, a quantitative risk characterization on L. monocytogenes in RTE foods; starting from the retail stage. EFSA Supporting Publications. 14(7). doi:10.2903/sp.efsa.2017.en-1252.

Pielaat A, Boer MP, Wijnands LM, van Hoek AHAM, Bouw E, Barker GC, Teunis PFM, Aarts HJM, Franz E. 2015. First step in using molecular data for microbial food safety risk assessment; hazard identification of Escherichia coli O157: H7 by coupling genomic data with in vitro adherence to human epithelial cells. International Journal of Food Microbiology. 213:130–138. doi:10.1016/j.ijfoodmicro.2015.04.009.

Possas A, Valdramidis V, García-Gimeno RM, Pérez-Rodríguez F. 2019. High hydrostatic pressure processing of sliced fermented sausages: A quantitative exposure assessment for Listeria monocytogenes. Innovative Food Science and Emerging Technologies. 52:406–419. doi:10.1016/j.ifset.2019.01.017.

Pouillot R, Goulet V, Delignette-Muller ML, Mahé A, Cornu M. 2009. Quantitative risk assessment of listeria monocytogenes in french cold-smoked Salmon: II. Risk characterization. Risk Analysis. 29(6):806–819. doi:10.1111/j.1539-6924.2008.01200.x.

Pouillot R, Hoelzer K, Chen Y, Dennis SB. 2015. Listeria monocytogenes Dose Response Revisited-Incorporating Adjustments for Variability in Strain Virulence and Host Susceptibility. Risk Analysis. 35(1):90–108. doi:10.1111/risa.12235.

Pouillot R, Miconnet N, Afchain AL, Delignette-Muller ML, Beaufort A, Rosso L, Denis JB, Cornu M. 2007. Quantitative risk assessment of Listeria monocytogenes in French cold-smoked salmon: I. Quantitative exposure assessment. Risk Analysis. 27(3):683–700. doi:10.1111/j.1539-6924.2007.00921.x. [accessed 2021 Jun 26]. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1539-6924.2007.00921.x.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. [accessed 2021 Jan 10]. https://www.R-project.org/.

Ray S. 2017. Support Vector Machine Algorithm in Machine Learning. Analytics Vidhya. [accessed 2021 Jun 15]. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

Ronholm J, Nasheri N, Petronella N, Pagotto F. 2016. Navigating microbiological food safety in the era of whole-genome sequencing. Clinical Microbiology Reviews. 29(4):837–857. doi:10.1128/CMR.00056-16. [accessed 2021 May 7]. https://pubmed.ncbi.nlm.nih.gov/27559074/.

Scholz M. Pangenome. Metagenomics. [accessed 2021 May 7]. http://www.metagenomics.wiki/pdf/definition/pangenome.

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 30(14):2068–2069. doi:10.1093/bioinformatics/btu153.

Sun P, Reid MD, Zhou J. 2014. An improved multiclass LogitBoost using Adaptive-One-Vs-One. Machine Learning. 97(3):295–326. doi:10.1007/s10994-014-5434-3. [accessed 2021 Jun 29]. https://link.springer.com/article/10.1007/s10994-014-5434-3.

Tirloni E, Nauta M, Vasconi M, di Pietro V, Bernardi C, Stella S. 2020. Growth of Listeria monocytogenes in ready-to-eat “shrimp cocktail”: Risk assessment and possible preventive interventions. International Journal of Food Microbiology. 334:108800. doi:10.1016/j.ijfoodmicro.2020.108800.

Tirloni E, Stella S, de Knegt L v., Gandolfi G, Bernardi C, Nauta MJ. 2018. A quantitative microbial risk assessment model for Listeria monocytogenes in RTE sandwiches. Microbial Risk Analysis. 9:11–21. doi:10.1016/j.mran.2018.04.003.

9. Appendix 1

QMRA studies on *L. monocytogenes*:

Author, year, country	QMRA method	Primary purpose of the study	Data Used	Main Findings	Remarks
Bemrah et al. 1998	Classic	Determine the expected number of listeriosis cases and deaths per year (for low and high-risk populations) due to soft cheese made from raw milk for a 50 million inhabitants' population and evaluate the consequences of farm-level interventions in the reduction of Listeria in milk and the decrease of consumption by populations at risk	French data from dairy farms on the origin of bovine raw milk contamination; Typical process for soft cheese making; Consumption per capita estimated from French data; Percentage of high-risk population	The average number of expected cases of listeriosis per year was 57 for a high-risk sub-population and 1 for a low-risk healthy subpopulation. The average number of expected deaths was 12 for the high-risk sub-population and 0 deaths for the low-risk sub-population. The alternative scenarios showed positive results.	Based on old data as recent data was confidential and could not be used; Infection by L. monocytogenes is associated with only a few virulent strains. Hence, this study multiplied the number of ingested bacterial cells by a factor in the range 0.01-0.1
Campagnollo et al. 2018a	Classic	Due to safety concerns arising from bad practices, determine the expected number of listeriosis cases in a population of 10 000 due to the consumption of two types of Brazilian cheese and evaluate the effects of changing the starting concentration of L. monocytogenes and the effects of the presence of anti-listerial lactic acid bacteria.	Listeria cells lost in milk coagulation from literature; Cheese yield from literature; Cheese weight from research, legislation and online survey; Challenge tests; Time and temperature during distribution to retail; Retail storage time and temperature; Brazilian domestic storage time and refrigerator temperatures; Data on cheese consumption	Consumption of semi-hard cheese would cause a mean of 26 cases of listeriosis. Fresh soft cheese consumption would result in 3443 and 4897 mean illnesses in general and vulnerable respectively. Scenario analyses indicated that aging of semi-hard cheese and inclusion of antimicrobial LAB mix in semi-hard and soft cheeses are effective risk mitigation measures.	The availability of more data could improve the results
Pouillot et al. 2007 and Pouillot et al. 2009	Classic	Due to listeria frequent isolation and its potential to grow in cold-smoked salmon, determine the annual number of invasive listeriosis cases from cold-smoked salmon consumption in	French samples of cold-smoked salmon; Monitorization data regarding time temperature; French data on consumption; Probability of invasive listeriosis from exposure to	Estimated to be 307, with a very large credible interval [10; 12 453], reflecting data uncertainty. This uncertainty is mainly associated with the dose-response model and the majority of	The assumption that the hazard characterization do not depend on the food considered is very common, however studies suggest an effect of foods on the virulence levels of

		France on three vulnerable populations (pregnant, susceptible, +65 years) plus reference and evaluate mitigation strategies	one cell based on literature and United States data; Data of invasive listeriosis cases from previous years;	cases are due to a very high level of contamination at the time of consumption linked to time-temperature abuse during the consumer step, rather than high initial levels therefore the best mitigation strategies are those concerning the consumer phase	L. monocytogenes strains. The results obtained from QMRA models should not be in contradiction with epidemiological data observed in the same area during the same period of time, if available.
Tirloni et al. 2018	Classic	Due to sandwiches intensive manipulation during production, and due to the use of different ingredients, estimate the expected number of listeriosis cases due to the consumption, on the last day of shelf life, of 20 000 servings of multi-ingredient sandwiches produced by a medium scale food producer in Italy, by different populations (healthy, susceptible, transplant recipients and total population) and evaluate different possible interventions	Food product samples; Challenge tests; Probability of invasive listeriosis from exposure to one cell based on literature; Percentage of transplant recipients based on Italian data;	0 cases were expected while 3 cases were expected when a higher variability in virulence and susceptibility was considered. The number of cases increased to 45–52 in the worst scenario (bean cream contamination and all transplant recipients consumers). Tested interventions resulted in a strong decrease of the risk but modified atmosphere packaging, should be regarded as the most promising one, as it can be performed by the producer	QMRA is more precise when a distribution of the probability of invasive listerioses from exposure to one cell is included.
Tirloni et al. 2020	Classic	The “shrimp cocktail” is a ready to eat product with the natural presence of the pathogen in the shrimps. This study estimates the expected number of invasive listeriosis cases caused by the consumption of 10 000 servings of the product on the last day of its shelf life,	Food product samples; Challenge tests; Percentage of susceptible people based on French data; Percentage of transplant recipients based on Italian data;	The model predicted 0 cases. The possibility of Listeria growth in the product could not be avoided. Treatment for 2 days in the lactic acid solution was selected as a good method to avoid listeria growth based on efficacy, the absence of consumer-perceptible sensorial modifications, and	

		considering a healthy, susceptible, and transplant recipients consumers. The exposure assessment model was also used to estimate the probability of the product exceeding the threshold of 2 log CFU/g during the shelf life. Modification of the production process was tested to re-classify the product as unsuitable for Listeria growth		the producers' production rate requirements.	
Brusa et al. 2021	Classic	Risk of listeriosis associated with the consumption of fermented sausages and dry cured pork shoulder contaminated with L. monocytogenes in Argentina and compare it to European standards	Data from butcher shops samples; Data on production process and time-temperature; Survey regarding consumption; Data and survey on susceptible population	The level of protection given by the current Argentine microbiological criterion (absence of L. monocytogenes in 25 g of product) would guarantee the safety of these products similarly to the <100 cfu/g cut used in other countries	
Possas et al. 2019	Classic	Build a probabilistic model to predict the growth of L. monocytogenes in Spanish chorizo sausage and verify the impact of high hydrostatic pressure (HHP) treatments on lowering microbial levels, and of changes in chorizo formulation, including lowering nitrite concentrations	Food product samples; Traditional processing; Challenge tests; Data on time and temperature based on manufacturers and literature	Cross-contamination during slicing was an important source contributing to increase pathogen prevalence and concentration. HHP is a powerful method for controlling L. monocytogenes in the final products. Healthier products, such as nitrite-reduced, obeying EU/US regulations for L. monocytogenes can be used associated with HHP	
Garrido et al. 2010	Classic	Develop risk assessment of Listeria in RTE products at a regional level,	Food product samples; Data regarding the population of Navarra;	The consumption of cooked ham is responsible for the higher prediction of cases.	

		since few risk assessment studies have been carried out using data from Spain. Identify the problems and data gaps that could be improved for future microbiological risk assessment and evaluation alternative scenarios	Population subgroups based on databases and literature; Data of consumption per year; Survey regarding serving size; Data on time–temperature of storage profiles before consumption; Optimal growth rate estimation based on database	Temperature storage at 4ºC through the food chain (including storage at home) has been demonstrated to be the best tool to decrease the risk of this serious illness. Informative campaigns to consumers are very important. There is the need for continuous studies on hygiene, storage and consumer behaviour	
Mataragas et al. 2010	Classic	Perform a risk assessment for deli meats and demonstrate how the QMRA can produce useful information for risk managers helping the development of intervention measures for reducing listeriosis cases	Data from literature and interviews with experts; Data on consumption;	The QMRA model can be used to evaluate the effectiveness of different risk management measures such as antimicrobials addition, thermal treatment of the final product, application of high hydrostatic pressure or irradiation, and decrease of product shelf life at or close to its threshold value	
Njage et al. 2018	WGS	Associate WGS with frequency of clinical cases since the limitations of existing risk assessment efforts are that dose–response models for L. monocytogenes do not always take into account differences in survivability and virulence among strains.	Isolates from food and clinical sampled in France; Literature review for all known virulence-associated factors, virulence factors, virulence genes, and environmental stress tolerance genes	The virulence genes FAM002725, FAM002728, FAM002729, InlF, InlJ, Inlk, lisY, lisD, lisX, lisH, lisB, lmo2026, and FAM003296 were important predictors of higher frequency of illness. These occur more frequently in ready-to-eat, dairy, and composite foods	Efforts have been made to account for virulence variability in L. monocytogenes but higher-resolution data such as WGS is needed. These models can reduce the number of unnecessary withdrawals of food with non or low-pathogenic strains and early detect the evolution of new pathogenic strains
Njage et al. 2020	WGS	Predict the potential for microbial growth and survival based on WGS, calculate the probability of	WGS data from food and food processing environment, from Canada and Switzerland and associated data on	Increased resistance to stress conditions leads to increased growth, the likelihood of higher exposure and probability of	An important benefit is the reduction in uncertainty in the exposure assessment models and while

		illness and the expected number of cases for three subpopulations including the healthy, susceptible and transplant recipients, and evaluate three alternative scenarios	growth phenotypes during different stress conditions, obtained from literature; WGS data from European Nucleotide Archive (ENA) under project number PRJEB15592; Microbial concentrations, initial contamination, consumer storage time and portions consumed per serving from literature.	illness. 0, 2 and 790 people ill in a population of 1 million people were predicted for healthy, susceptible and transplant recipients respectively	making risk estimates. Neglecting within-species heterogeneity in microbial stress response may compromise the QMRA quality and the reliability of evidence used for control efforts.
Fritsch et al. 2018	WGS	Show how recently published genomic data can be used to refine a model assessing the listeriosis risk linked to the consumption of cold-smoked salmon as nowadays lack of accurate information on the variability of the dose-response for foodborne pathogens induces a significant uncertainty in risk estimates	WGS data from France; WGS data from smoked salmon in Europe; WGS data from food and food processing environment, from Canada and Switzerland and associated data on growth phenotypes during different stress conditions, obtained from literature;	The generic model in which the population structure in cold-smoked salmon was not considered, predicted a number of 978 listeriosis cases while in the model taking into accounts the prevalence of the subpopulations in cold-smoked salmon, 574 listeriosis cases were predicted.	This might improve the hazard identification, risk estimation and risk management *e.g*. by adapting or by refining the intervention strategies according the hazards' properties
Fritsch et al. 2019	WGS	Get better insight into different existing bioinformatics approaches to associate bacterial phenotype(s) and genotype(s)	WGS data from multiple sources; Experiments to determine phenotypic variability of growth at low temperature;	Some genes and SNPs were associated to growth at 2ºC	The successful application of combined bioinformatics approaches associating WGS-genotypes and specific phenotypes, could contribute to improve prediction of microbial behaviors in food.

Other WGS studies:

Author, year, country	Pathogen	Primary purpose of the study	Data Used	Main Findings	Remarks
Chen et al. 2020	Salmonella	Explore the applicability of WGS and other genomic technologies in Salmonella QMRA, pool genomic data associated with phenotypic data, and outline the added value and current limitations of the WGS usage	Summary of databases and initiatives for salmonella genomic analysis; Summary of comparative genomic studies of salmonella by WGS; Summary of comparative multiple virulence factor studies of Salmonella by PCR; Summary of comparative genomic studies of Salmonella by Comparative genomic hybridization with or without other methods; Summary of comprehensive proteomic or genoproteomic studies of Salmonella	Gaps and inconsistencies due to the limited number of related studies, diversity in the methodology of phenotypic testing, and variability in data obtained. Therefore, a quantitative analysis at this point might not be feasible.	The transition to genomic technologies is being done without a clearly defined path and methodology but the normalization of proteomics and the increase in the use of machine learning and data analytic techniques can improve the efficiency of QMRA in the future.
Pielaat et al. 2015	E.coli	Introduce a method for hazard identification linking WGS data with in vitro results	WGS data of Shiga toxin-producing Escherichia coli (STEC) O157; Phenotypic data (in vitro adherence to epithelial cells as a proxy for virulence)	This application revealed practical implications when using SNP data for QMRA. A new challenge is the translation of multidimensional genotypic information to a single measure of risk without losing relevant information	With WGS data the cost and time shifts from data acquisition to data analysis.
Njage et al. 2019	E.coli	Explore the potential of machine learning in the prediction of illness outcome resulting from shigatoxigenic E. coli (STEC) infection.	STEC isolates collected over 5 years where patients were interviewed for information including severity of infection, travel history, epidemiological linkage and specific source of infection	The most important predictor protein families are A0747, A0253, A0259, A5715, A2240, A0434, A0702, A0710, A0712, A0882, A0899, A0925, A0942, A3466, A3764, A4831, A4856, A0508, A0898, A0932 and A0960.	The application of machine learning utility will continue to rise as more WGS data accompanied by clinical outcome becomes available and will allow the detection of new threats reducing reaction time during an outbreak and avoiding unnecessary product withdraws.

10. Appendix 2

Cheese Type	Number of Samples
Aged hard cheese	1
Aged raw milk cheese	4
American curd cheese	1
Asadero cheese	2
Blue cheese	6
Blue cheese crumbles	1
Bovine cheese	1
Brie cheese	2
Bucheron goat cheese	2
Burrata soft cheese	3
Camembert goat cheese	1
Cheddar cheese ball	2
Cheese	138
Cheese curd	1
Cheese from sheep milk	1
Cheese spread	1
Chihuahua type cheese	1
Cotija cheese	4
Cow/sheep milk cheese	2
Cream cheese	1
Cubed cheddar cheese	2
Fermier goat cheese	2
Fontina cheese	5
Fresh cheese	12
Fresh cheese curd	18
Fresh mexican style cheese	2
Fresh round cheese	1
Fresh white cheese	1
Goat cheese	3
Gorgonzola cheese	1
Gorgonzola dolce cheese	1
Gouda cheese wheel	1
Hard cheese	2
Home-made cheese	1
Le vigneron marc cheese	1
Mexican cheese	2
Mexican soft cheese	7
Mexican white cheese	2
Mold-ripened blue-veined cheese	1
Moliterno al tartufo cheese	1
Morbier cheese (aged over 60 days)	1
Mozarella cheese	1
Oaxaca string cheese	1
Pasteurized cows' milk cheese	1

Pasteurized milk queso fresco cheese wheels in vac-packed plastic	1
Pasteurized milk ripened semi-soft mozzarella (bocconcini)	1
Quesillo oaxaca string cheese	2
Queso cotija	1
Queso fresco cotija	1
Queso seco cheese	1
R. Salinas cheese	1
Raw milk cheese	13
Raw milk cheese aged 60 days	1
Raw milk cheese-monterey jack	1
Ricotta cheese	7
Ricotta piatta cheese	1
Ripened pasteurized milk semi-soft ricotta cheese	1
Ripened pasteurized milk soft brie (camembert) cheese	1
Ripened pasteurized milk soft cheese curds	1
Ripened pasteurized milk soft cream cheese	1
Robiola pineta cheese	3
Salvadorian string cheese	1
Semi soft cheese	4
Sheeps milk cheese	3
Sheep's milk ricotta cheese	1
Shredded cheddar cheese	1
Shredded mozzarella cheese	1
Soft cheese	6
Soft ripened cheese	4
Soft white mexican cheese	2
Spanish cheese	2
Spreadable cheese	1
Swiss cheese	1
Talleggio cheese	2
White cheese	6