Escola Nacional de Saúde Pública

UNIVERSIDADE NOVA DE LISBOA

Large-scale investigation of a COVID-19 high-incidence region of Portugal: an outbreak analysis using Genomic Epidemiology

Master's in Public Health

Ana Regina Serra Sá

November 2021

Escola Nacional de Saúde Pública

UNIVERSIDADE NOVA DE LISBOA

Large-scale investigation of a COVID-19 high-incidence region of Portugal: an outbreak analysis using Genomic Epidemiology

Dissertation presented to meet the requirements necessary to obtain the master's degree in Public Health carried out under the scientific guidance of Andreia Leite, PhD, National School of Public Health (ENSP-UNL) and Vítor Borges, PhD, National Institute of Health Dr. Ricardo Jorge (INSA).

November 2021

Acknowledgements

I would like to offer my special thanks to my advisor, Andreia Leite, from the Nova National School of Public Health (ENSP-UNL), who guided me through this process, patiently and wisely. Her guidance was essential to give me focus and helped me gain essential tools for the research process.

I would like to express my sincere gratitude to all the team from the National Institute of Health (INSA), especially my co-advisor, Vitor Borges, but also Joana Isidro and João Paulo Gomes, who extensively contributed to this research.

I am deeply grateful to Judite Matias, my medical residency tutor, and Ana Oliveira, my coordinator, from the Public Health Unit of Baixo Vouga, for all the support in developing this project and for believing in me to deliver quality research with added value for local Public Health decision.

I would like to extend my sincere thanks to Sofia Tedim, University of Aveiro, and Baltazar Nunes, Nova National School of Public Health (ENSP-UNL), for their insightful contributions to the statistical analysis. Additionally, Constantino Caetano, INSA, for the technical help in solving script-related problems.

Abstract

Background

Local outbreak response is highly dependent on the epidemiologic investigation (EI), which relies on subjective information to reconstruct transmission chains. This study aims to describe COVID-19 transmission chains in the light of viral genomics. Additionally, it aims to estimate the sensitivity of the EI.

Methods

A combination of epidemiological and viral genomic data was performed to understand COVID-19 transmission chains, in Baixo Vouga, a region with a municipality in sanitary cordon and presenting one of the highest incidence rates during the first epidemic wave. Data previously collected by the public health authorities was used to reconstruct epidemiological clusters (epiclusters). The genome sequences were retrieved from the nationwide SARS-CoV-2 collection. Sensitivity was determined as the proportion of cases in each epicluster that belonged to the same main viral genetic profile. A metanalysis of epicluster-specific sensitivity was conducted.

Findings

A total of 1925 confirmed cases of COVID-19 were identified, from 8th March to 30th June 2020. From these, 128 cases had viral genomic data available. Two large epiclusters were identified (280 and 101 cases), presenting a central role on the spread of the disease. Genomics was also key to resolve some sporadic cases and misidentified direction of transmission. The sensitivity of epidemiological investigation to detect transmission chains was 80% (95% confidence interval: 70%-86%).

Interpretation

Epidemiological reconstruction of the transmission chains presented high sensitivity. However, high incidence and closed settings led to misidentification of links. This study contributes to understand the hurdles and caveats associated with the EI in the context of a massive outbreak of COVID-19.

Keywords: SARS-CoV-2; COVID-19; epidemiology; pathogen genomics; transmission chains.

Resumo

Introdução

A resposta local aos surtos é altamente dependente da investigação epidemiológica (IE), que, por sua vez, se baseia em informações subjetivas para reconstruir as cadeias de transmissão. Este estudo tem o objetivo de descrever as cadeias de transmissão de COVID-19 à luz da genómica do vírus. Adicionalmente, pretende-se estimar a sensibilidade da IE.

Metodologia

Foram combinados dados genómicos e epidemiológicos para compreender as cadeias de transmissão de COVID-19, no Baixo Vouga, uma região com um município em cordão sanitário e com uma das incidências mais elevadas da primeira onda em Portugal. Dados previamente colhidos pelas autoridades de saúde foram utilizados para a reconstrução dos clusters epidemiológicos (epiclusters). As sequências do genoma viral foram recuperadas da coleção nacional de SARS-CoV-2. A sensibilidade foi determinada como a proporção de casos, em cada epicluster, que pertencia ao mesmo perfil genético viral. Foi conduzida uma metanálise para determinar a sensibilidade combinada.

Resultados

Foram identificados 1925 casos confirmados de COVID-19, entre 8 de março a 30 de junho de 2020. Destes, 128 casos tinham dados genómicos virais disponíveis. Foram identificados dois grandes epiclusters (280 e 101 casos), que apresentavam um papel central na propagação da doença. A genómica foi fundamental para esclarecer alguns casos esporádicos e direções de transmissão mal identificadas. A sensibilidade da IE para detetar cadeias de transmissão foi de 80%, com um intervalo de confiança a 95% entre 70% e 86%.

Discussão

A reconstrução epidemiológica das cadeias de transmissão apresentava alta sensibilidade. No entanto, situações de alta incidência e contextos fechados conduziam a um erro na identificação das ligações epidemiológicas. Este estudo contribui para a compreensão dos obstáculos e equívocos associados à IE no contexto de um surto de COVID-19.

Palavras-chave: SARS-CoV-2; COVID-19; epidemiologia; genómica de patógenos; cadeias de transmissão.

Table of contents

List of tables

List of figures

Abbreviations

Baixo Vouga Primary Care Cluster (BVPCC)
Basic Reproductive Number (R_0)
Confidence Interval at 95% (95%CI)
Disability-Adjusted Life Years (DALYs)
Effective Reproductive Number (R_e)
Epidemiologic Investigation (EI)
Epidemiological Investigation (EI)
Hierarchical Codes (HC)
Human Immunodeficiency Virus (HIV)
Identification Number (ID)
Interleukin-6 (IL-6)
International Health Regulations (IHR)
Interquartile Range (IQR)
Middle East Respiratory Syndrome coronavirus (MERS-CoV)
National Epidemiological Surveillance System (SINAVE)
National Institute of Health (INSA)
Novel Coronavirus Disease (COVID-19)
Portuguese Directorate-General of Health (DGS)
Primary Care Cluster of Baixo Vouga (PCCBV)
Public Health Emergency of International Concern (PHEIC)
Public Health Unit (PHU)
Public Health Unit of the Primary Care Cluster of Baixo Vouga (PHUPCCBV)
Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR)
Severe Acute Respiratory Syndrome (SARS)
Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)
Single Nucleotide Polymorphism (SNP)
Social-Demographic Index (SDI)
Whole-Genome Sequencing (WGS)
World Health Organization (WHO)
Years Lived With Disability (YLDs)

1. Introduction

The novel coronavirus disease (COVID-19) was first reported in patients with atypical pneumonia, in December 2019, in China (1,2). The COVID-19 outbreak was declared a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and a pandemic by the World Health Organization (WHO) on March 11, 2020 (3,4). Europe was, for several weeks, the most severely affected continent. Portugal had its first detected cases of the disease reported on March 2, 2020, reaching 42 171 cases and 1 576 deaths by 30th June, 2020 (5). In Portugal, the most affected regions in the first wave were the North and Centre. Baixo Vouga Region (NUTS III), with a total of 370 394 inhabitants in 2011 and 11 municipalities, belongs to the centre (NUTS II) and borders the North region, presenting as a buffer towards south. The first case in Baixo Vouga was on March 8th, in Aveiro municipality (6). There was a total of 1925 confirmed COVID-19 cases from March 8th to the end of June in Baixo Vouga, 39% belonging to one municipality, Ovar (6). With about 54 120 inhabitants in 2018, Ovar was the first Portuguese municipality to be declared to have active community transmission (7). On the same day, a sanitary cordon was decreed in this municipality, lasting between March 17 and April 17, 2020 (8).

Local outbreak response is highly dependent on the epidemiologic investigation. However, epidemiological investigation relies on subjective information to reconstruct transmission chains and cannot guarantee total reliability. Thus, viral whole-genome sequencing (WGS) facilitates the identification of phylogenetic clusters of SARS-CoV-2, allowing the surveillance of the viral lineages in circulation (9,10), the occurrence of introduction events (11–20) and the reconstruction of transmission chains by informing the putative source of transmission (21–30). To date, few studies have scrutinized the concordance between viral genomics and epidemiological data on the reconstruction of COVID-19 transmission chains, and only one provided a quantitative estimate (22).

Here, a combination of epidemiological and viral genomic data is performed to understand COVID-19 transmission chains, from March to June 2020, in Baixo Vouga Region, an area with a municipality with a sanitary cordon and presenting as one of the regions with higher incidence rates in the first wave. The main objective of this study was to assess the concordance of links as defined by epidemiologic investigation and genome sequencing of SARS-CoV-2, and further describe the context of the misidentifications of the epidemiologic links. By taking advantage of viral genomic data, this study expects to contribute to the understanding of the hurdles and caveats

associated with the epidemiological investigation of hundreds of community cases in the context of a massive outbreak caused by a highly transmissible and new respiratory virus.

This thesis is also a result of the genomic surveillance of SARS-CoV-2 implemented by the National Institute of Health Dr. Ricardo Jorge (INSA) in collaboration with a nationwide consortium of >50 hospitals/laboratories. All the process from sample collection to phylogenetic analysis was performed by INSA. For the purpose of this project a phylogenetic tree, hierarchical codes and main viral genetic profiles were produced by INSA and then linked to the study population of Baixo Vouga in order to produce a common database with both epidemiological and viral genomic data. The process of epidemiological database treatment and linkage with genomic data was performed by me as well as the following analysis presented. Interpretation of results of this combined data integration was performed in close collaboration with the INSA team. As part of this project a paper has been submitted for publication.

This document starts by describing main concepts and epidemiology of communicable diseases, stating the problems associated to them and underlining its importance (Communicable diseases: main concepts and epidemiology). Then, it presents a description of the novel coronavirus disease (COVID-19) in terms of its epidemiological evolution, clinical characteristics, transmission, diagnosis and management, with a special focus on non-pharmacological interventions (The novel coronavirus disease (COVID-19)). It then present the main concepts of epidemiological surveillance and the specific context of Portugal and Baixo Vouga (Epidemiological Surveillance). Lastly, it introduces the use of Genomic Epidemiology to resolve issues when considering genomic and epidemiology separately to prevent and control the dissemination of infectious diseases (Public Health Genomics). In the second section, it presents the study design, setting, and data sources for epidemiological data, alongside with the specific information on genomic sequencing and bioinformatics analysis (Methods). It then presents the statistical analysis, including an in-depth explanation of the decision to estimate epidemiologic investigation sensitivities, at the epicluster level and combined. In the third section, it presents the results, starting with a contextual description, then study population, sample and epiclusters (Results). This section provides detailed information on the viral genomic diversity found across epiclusters and on the epidemiological investigation within large epiclusters. Then, is presented a description of the inputs of genomics to resolve sporadic cases and the misidentifications of epidemiological links directions. In the end of this section, a presentation of the main

analysis is presented, starting with epicluster-specific sensitivities of the epidemiological information and then the pooled estimate. After, a discussion of the results confronts them with the previous literature and present the main strengths and limitations, as well as implications and added values of this research (Discussion). To end, this document presents the main conclusions and recommendations (Conclusions).

2. Background

a) Communicable diseases: main concepts and epidemiology

Infectious diseases are a major cause of morbidity, disability, and mortality worldwide. An infection occurs when a capable agent enters a human or animal body and develops or multiplies. Infection can be unapparent (named as colonization or carrier status) or result in manifestations due to tissue function impairment (disease) (31). Infectious diseases thus refer to all diseases caused by microorganisms (a specific agent or its products) while communicable diseases refer to those that can be transmitted from an infected person to another, directly or indirectly (32). Not all infectious diseases are communicable. For example, tetanus is not communicable as it is the result of the human contact with contaminated soil (33). Agents causing infectious diseases include viruses, bacteria, fungi, parasites, and prions. Transmission of an agent or its products occurs from an infected person, animal, or reservoir to a susceptible host (31), directly or indirectly. Direct routes include mucous to mucous membrane contact (e.g. sexually transmitted diseases), droplet spread by sneeze or cough (e.g. influenza), vertical across placenta (e.g. toxoplasmosis), trough organ transplant or blood transfusion (e.g. hepatitis B), or from skin to skin (e.g. herpes simplex type 1) (31,32). Indirect transmission can occur through an intermediate plant or animal host (vector-borne), the inanimate environment (vehicle-borne), or microbial aerosols, like droplet nuclei or dust (airborne) (31,34,35). The natural habitats in which infectious agents live and multiply so that they can be transmitted to a susceptible host are called reservoirs (31). These can be humans, animal, arthropod, plant, soil, or substance, or a combination of these, as for example tepid water for *Legionella*. Reservoirs differ from the source, which is the actual origin from which the infection is acquired (32). Several models have been proposed to explain disease occurrence, from which the epidemiologic triad (Figure 1) is the most traditional one. It acknowledges three main aspects in the causation of disease: the external agent, a susceptible host, and an environment that makes possible the encounter between host and agent (36). These factors' balance and their interaction differ according to the specific disease. For infectious diseases, transmission occurs when the agent leaves its reservoir or host through a portal of exit, is conveyed by some mode of transmission, and enters through an appropriate portal of entry to infect a susceptible host (36). This sequence is called the chain of transmission (37). Understanding the occurrence of disease requires a knowledge of each step in the chain of transmission, including factors that influence the exposure to the infectious agents and

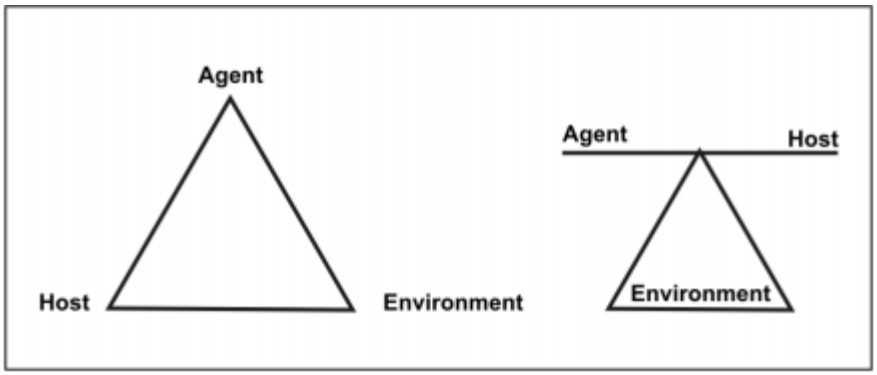

Figure 1 Epidemiologic triad. Source: Centre for Disease Prevention and Control(5)

the transmission itself. Exposure will depend on the specific agent, its modes of transmission and the probability and frequency of a contact between infected and susceptible. It is important to distinguish between exposure to a reservoir and exposure to the infectious agent since reservoirs may or may not represent a risk of exposure to the agent. Once there is an exposure, transmission depends on infectiousness, and the proportion of susceptible individuals. Infectiousness varies according to the infectious agent, and it is commonly highest just before the onset of disease (pre-symptomatic phase) and ends before disease resolution (37). In addition to the characteristics of the agent, host immunity and antimicrobial treatment, disease resolution also influences infectiousness (Figure 2) (37). These factors can be translated by numerical indicators, the basic and effective reproductive number (32,37). The Basic Reproductive Number (R_0) corresponds to the average number of secondary cases in a completely susceptible population following the introduction of a single infectious case. The Effective Reproductive Number (R_e) is the average number of transmissions of infection that occurs from each infectious case (e.g., the average number of secondary cases from each primary case) (37). The process of infection follows the exposure. This is dose-related, agent-related and host-related. The time from infection to disease occurrence varies widely and is called incubation period. It can be influenced by certain risk factors as age or underlying medical conditions. When an agent is transmitted, host and agent characteristics will also determine if the disease develops.

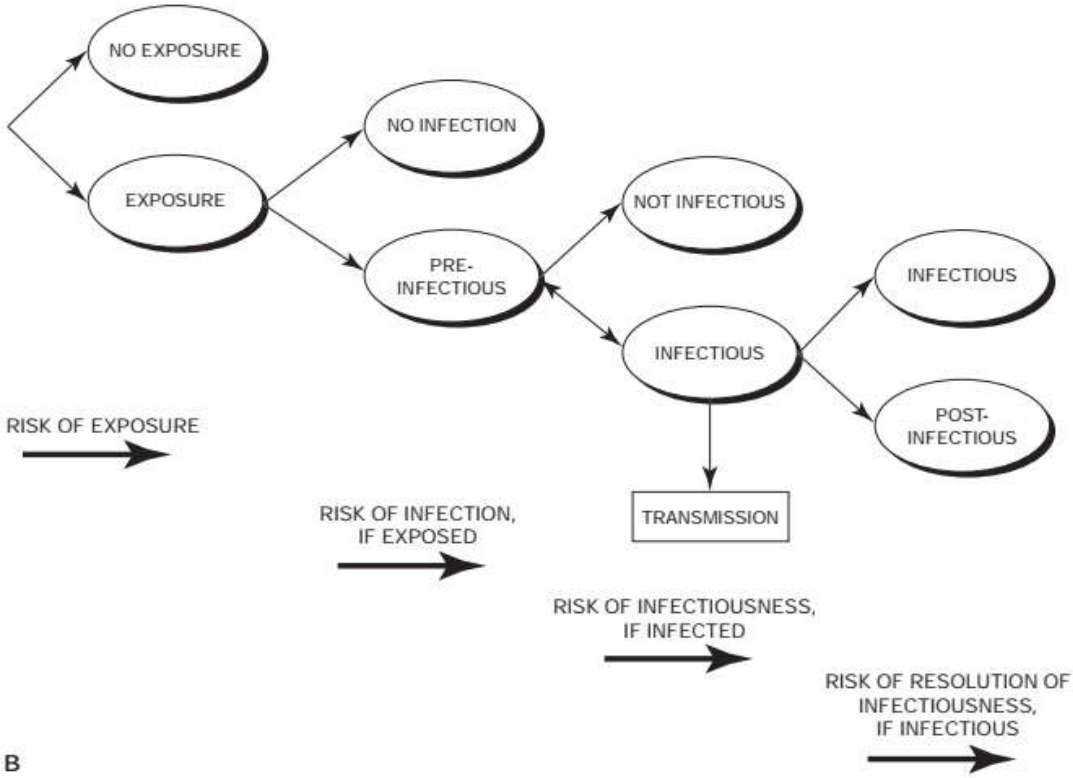

Figure 2 Schematic time course of states of transmission potential from exposure to post infectiousness. Source: Rothman et al. (37)

The factors implicated in the transmission chain that can influence the probability of development and spread of communicable diseases are called determinants. They are related to the elements of the epidemiologic triad and that can be, for example, socioeconomic, environmental, and behavioural factors, as well as international travel and migration. Determinants associated with the agent include, for example, its pathogenicity (the capacity of producing disease), virulence (the degree of pathogenicity, a concept of disease severity) or the ability to survive antibiotics (antimicrobial resistance) (31,36). Host intrinsic factors also influence the individual's exposure, susceptibility, or response to a causative agent (36). For example, behavioural determinants of the Human Immunodeficiency Virus (HIV) include unsafe sex (responsible for 76% of total Disability-Adjusted Life Years (DALYs) HIV-related) and the use of illegal drugs (8% of total DALYs) (38). Finally, extrinsic, environmentally related factors, affect the agent and the opportunity for exposure, and can be physical, biological, chemical, socioeconomic, among others. Crowding is a very well known risk factor for various communicable diseases, either in community settings, potentiating respiratory tract diseases, meningitis, typhus, and scabies, among others, or in healthcare settings, increasing the risk of hospital-acquired infections (39). In 2019, unsafe water, poor sanitation, and insufficient handwashing were the attributed risk factors of 76% of the

total DALYs of enteric infections (38).

i. Epidemiological transition

Over time, and in particular during the last 150 years, the improvement of certain environmental determinants and the increasing effectiveness in treatment, prevention, and control, led to a decrease in morbidity and mortality due to infectious diseases (40). From 1990 to 2019, all the countries in the world, except for Ukraine and Lesotho, saw a decrease in the ratio of Years Lived With Disability (YLDs) from communicable, maternal, neonatal and nutritional diseases to the non-communicable and injuries (41). These trends are part of a phenomenon of reduction of communicable diseases and an increase in non-communicable ones, the so-called epidemiological transition.

The epidemiological transition comprises a complex course of development of determinants (demographic, socioeconomic, cultural, biological, technological, and environmental) which results in modifications of the mortality patterns (40). The theory of epidemiological transition as described by Omran (42) has three major stages: (1) the age of pestilence and famine, (2) the age of receding pandemics, and (3) the age of degenerative and man-made diseases. For European countries, the 1800's were mainly marked by the consequences of war and natural disasters, and with great expression of diseases like typhus, cholera, or smallpox (43). The provision of sewers and waterworks allowed better hygiene and a reduction of the death rates due to diarrheal diseases (44). The advent of industrialisation has concomitantly allowed better nutrition patterns and the reduction of famine. Scientific advances, including the Louis Pasteur theory of germs (1857) and the subsequent advent of vaccination (developed by Edward Jenner in 1796 for smallpox) played a particularly important role in the reduction of infectious diseases. Vaccination made the complete eradication of one communicable disease possible and highly reduced the incidence of others. Smallpox was eradicated worldwide in 1979 and efforts are still being made by the international community to eradicate poliomyelitis (45). The early XX century, not only presented with additional important technological advances, like the discovery of penicillin, in 1928 by Alexander Fleming, but was also the moment in which developed countries introduced systematic vaccination.

Thanks to the combination of these previous factors that lead to epidemiological transition, in 1990, the top 5 mortality causes, worldwide, included 3 communicable diseases, while in 2019, only one was in this list (38). With increasing Social-Demographic Index (SDI), a composed measure that allows the positioning of counties in the spectrum of development, a shift in the burden of disease from communicable, maternal, neonatal, and nutritional diseases towards non-communicable and injuries is

expected (41). In 2019, communicable diseases represented only 13.9% of all deaths worldwide (38). The top 5 major causes of mortality, globally, in all ages, in the year of 2019, included only one, in the fourth position, which was of communicable nature – respiratory infections and tuberculosis, with 47.6 deaths per 100 000 (38). Globally, seven out of the ten most important contributors to declining burden were of communicable nature (e.g., causes that had the largest absolute decreases in number of DALYs between 1990 and 2019)) (41). They were mostly diseases that predominantly affect children, such as lower respiratory infections, diarrhoeal diseases, neonatal disorders, measles, tetanus, and malaria (41), but also, one disease, which largely affects adults, tuberculosis (41) By avoiding premature dead by communicable diseases, high-income countries' population is more prone to suffer from other diseases that are typically associated with older age and certain risk factors as unhealthy diet, physical inactivity, sedentarism, obesity, hypertension, and tobacco use. This causes a switch towards a larger fraction of the burden in morbidity (as measured by YLDs) compared with early mortality (as represented by years of life lost (YLLs)) (41). However, middle-income countries present with a particular pattern of double-burden. There, the overall proportion of deaths from communicable diseases is decreasing, but they still remain a major public health challenge in these countries (40), explained by the gap seen in these countries' population living standards (41). The last phase of the epidemiological transition previews delayed degenerative diseases and emerging infections (42). This phase suggests, not only a postponement of degenerative diseases to older ages, but also the emergence of new and the resurgence of previously contained infectious diseases.

Despite the great reduction on communicable diseases worldwide, several challenges lie ahead (40), including the role of infectious agents in chronic diseases, antimicrobial resistance, climate change and (re)emerging infectious diseases. Firstly, several chronic diseases are directly or indirectly caused by infectious agents, including *Helicobacter pylori* which can lead to peptic ulcer disease and gastric carcinoma, or human papillomavirus, the causal agent of virtually all cervical cancer cases. More recently the role of commensal microorganisms (microbiota) has been acknowledge to influence humans in a normal healthy symbiotic relation (46). This is a field under development but a connection with chronic diseases seems plausible (47). Secondly, antimicrobial resistance is one of the biggest threats to global health. Misuse of antibiotics in humans, animals, and agriculture, is thought to be the main accelerator of a process that occurs naturally as a form of bacterial evolution (48). This phenomenon makes it harder for

clinicians to treat infections using the common antibiotics, leading to longer hospital stays, higher medication costs, and increased morbidity, mortality, and disability (48). One of the most worrying examples is the resistance of the *Mycobacterium tuberculosis*, that in 2017 accounted for around 600 000 cases of disease resistant to rifampicin – the most effective first-line drug (49). Thirdly, mosquito-borne diseases, like dengue or malaria, are still present in tropical settings but also expanding to new geographies due to climate change and expansion of the vector installation (49). Another field of concern regarding communicable diseases is bioterrorism, which corresponds to the weaponization of infectious agents. Finally, emerging and re-emerging infectious diseases have a great impact in global health and are a raising concern. High-income countries have been witnessing a phenomenon of resurgence in certain infectious diseases due to vaccination hesitancy, where the most iconic example is measles, with a 30% increase in cases globally (49). Ebola virus and other high-threat pathogens, present a critical challenge in a globalised world, where travelling is fast and massive. For this reason, a global public health effort towards international disease surveillance and control is critical, especially if effective treatment and vaccines are not yet available for certain diseases. This also includes Zika, Middle East Respiratory Syndrome coronavirus (MERS-CoV), Severe Acute Respiratory Syndrome (SARS), and now the pandemic due to the Novel Coronavirus Disease (COVID-19) (49), which is the focus of the current work.

ii. Communicable diseases epidemics

Epidemic is the occurrence, often sudden, in a certain population, of cases of an illness, specific health-related behaviours, or other health-related events clearly in excess of normal expectancy (31). The term refers both to communicable and non-communicable diseases. Outbreak is used to describe the localized occurrence of two or more related cases of a disease, or an increase in the observed incidence of cases over the expected incidence within a given time period and a restricted geographic area, or a single case of a serious disease (35). An outbreak is often described as a localized form of epidemic. The occurrence of outbreaks is a distinctive feature of many communicable diseases (37). A pandemic is an epidemic affecting multiple countries or continents and that usually involves a large number of people (36). Generally, a cluster is a small group (two or more) of people, things, or events. In epidemiology, this word is used to describe the aggregation, in space or time, of events in frequencies that are believed to be greater than expected (31,37). Eventually, a cluster can become an outbreak when the current frequency of disease is shown to be higher than the expected one. These previous terms

describe abnormal increases in diseases (or events) frequencies. To refer to the normal (baseline) level of a particular disease in a certain population the term endemic is used (31). Often, the latter does not refer to the desired level (which can be zero) but to the observed one and it can also be used to refer to the constant presence of a disease or infectious agent in a population (31,36). Hyperendemic describes persistently high levels of disease (36).

New diseases commonly emerge in human/micro-organism interactions, and old diseases can re-emerge in the form of antimicrobial resistance or due to vaccine hesitancy or bioterrorism. Also, the susceptibility and/or exposure of people to communicable diseases change, posing a constant challenge to health systems. Therefore, is of absolute priority to invest in the prevention of epidemics. This can be done through primary prevention, according to the routes of transmission. For diseases that can be transmitted directly, it can be done through the interruption of the chain of transmission, reduction of the susceptibility of the host and vaccination campaigns. For diseases of indirect transmission, it can be done through environmental protection, health education and community participation to promote vector control, environmental hygiene and food safety (35). Secondary prevention can be achieved with screening and outbreak investigation, in order to identify the cause, route of transmission, and risk factors, so that control and prevention strategies can be implemented (35).

b) The novel coronavirus disease (COVID-19)

The novel coronavirus disease (COVID-19) was first reported in patients with an atypical pneumonia, in December 2019, in China, epidemiologically linked with the Wuhan market (1,2). The disease is caused by The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).

SARS-CoV-2 is a RNA virus which encodes its genomic material on a single-stranded molecule, with a relatively small size genome around 30 000 nucleotides (50). SARS-CoV-2 shows a substitution rate (i.e., a measure of mutation accumulation in a given period of time and embeds the effects of selection (51)) of around 8×10^{-4} mutations per nucleotide site per year, which is equivalent to around 24 mutations per genome per year (or ~2 mutations per month) (50).

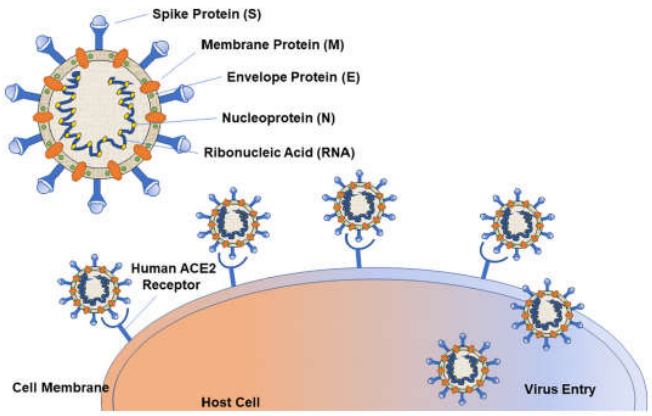

Figure 3 . Schematic representation of the SARS-CoV-2 structure and its mode of host entry. Source: Naqvia et al. (81)

SARS-CoV-2 has four structural proteins encoded in 11 genes: spike (**S**), envelope (**E**), membrane (**M**), and nucleocapsid (**N**) proteins (Figure 3). The N protein surrounds the genome of the virus and the S, E and M form the viral capsid. The spike protein (S) allows the virus to attach to the cellular membrane of the host and has been specially studied for vaccine development.

i. Epidemiological evolution

World, Europe, and Portugal

The first three cases detected in Europe were reported in France on January 24, 2020 (52). On January 30 2020 the outbreak was declared a Public Health Emergency of International Concern (PHEIC), which is a formal declaration by the World Health Organization (WHO) of an extraordinary event that constitutes a risk for other countries through cross-borders dissemination and that requires an international response (53). The situation was declared a pandemic by the WHO on March 11, 2020 (3,4). Almost every European and American country experienced the first wave of the epidemic in the spring and the second wave in the fall of 2020. As of December 27, 2020, a total of 79 231 893 cases of COVID-19 and 1 754 574 deaths have been reported by the WHO, worldwide (54). Europe was, for several weeks, the most severely affected region. Portugal had its first detected case reported on March 2, 2020, reaching 394 573 cases and 6 619 deaths by December 27, 2020 (55).

Baixo Vouga Primary Care Cluster (BVPCC)

The most affected Portuguese region in the first wave was the north. Baixo Vouga, with a total of 370394 inhabitants in 2011, and 11 municipalities, is located in the centre region, but it is immediately contiguous to the north (7). It is thus presented as a buffer for the centre region of the country. The first case in Baixo Vouga was on March 8, in Aveiro municipality. There was a total of 1 925 new cases detected of COVID-19 from March 8th to the end of June in Baixo Vouga, 38,9% belonging to one municipality, Ovar. With about 54 120 inhabitants in 2018, Ovar was the first Portuguese municipality to be

declared as having active community transmission (7). In the same day, a sanitary cordon was decreed, lasting between March 17 and April 17, 2020 (8). During this period only few authorized workers and ambulances could cross the controlled borders, and all services but the industry deemed essential were shut down. Unlike the rest of the country, where people depended on central validation from the Portuguese Directorate-General of Health (DGS) to have a test for SARS-CoV-2 approved, Ovar, due to the declared state of calamity (56,57), had an independent circuit that allowed more independence on the testing criteria. This context provided conditions for large-scale testing strategy at an early stage. The present study will focus on Baixo Vouga area, for the temporal period between March and June of 2020.

ii. Clinical characteristics

COVID-19 ranges from an asymptomatic infection to a severe disease. The disease presentation involves a broad spectrum of symptoms, including fever, cough (either with or without sputum), dyspnoea, fatigue, myalgia, arthralgia, sore throat, runny nose, headache, diarrhoea, nausea, vomiting and abdominal pain (58–61). Dermatological, and neurological symptoms have also been described. These include polymorphic lesions, erythema, chilblain-like and urticaria for the former and olfactory and gustatory dysfunction, like anosmia and ageusia, for the latter (60,62,62–69).

The virulence of COVID-19 can be seen through the case fatality ratio (CFR). A recent meta-analysis, showed an overall pooled CFR of 10.0% (95% CI: 8.0-11.0) for this disease (70). Hospitalized patients presented higher risk of death (13.0% (95% CI: 9.0-17.0)) compared to non-hospitalized (1.0% (95% CI: 1.0-3.0), and being admitted in the Intensive Care Unit (ICU) presented a CFR of 37.0% (95% CI: 24.0-51.0) (70). Older patients (over 50 years old) presented a CFR of 19.0% (95% CI: 13.0-24.0) (70). Besides age and clinical status, other risk factors have been associated with higher risks of deaths. Some comorbidities presented high Hazard Ratio (HR) or Odds Ratio (OR) associated with fatal COVID-19, as diabetes (HR 1.2-2.0), obesity (OR 1.5-1.75), heart failure (HR 1.3-3.), Chronic Obstructive Pulmonary Disease (HR 1.12-2.2), dementia (HR 1.4-7.7), liver cirrhosis (OR 3.2-5.9) and active cancer (OR 1.6-4.7) (71).

iii. Transmission

As explained in “Communicable diseases: main concepts and epidemiology”, transmission is determined by various factors such as the competence of the virus to replicate, the patient’s symptoms (e.g. cough that releases more infectious droplets), and the behaviour and environmental factors associated with the infected individual (72,73).

According to the existing evidence, SARS-CoV-2 can be transmitted directly or indirectly. Direct transmission occurs by person-to-person contact (through infected secretions) or through droplets. Indirect transmission might be airborne, through contact with fomites, or through faeces' contaminated environments (74–77). The most frequent transmission mode is directly, through droplets (78). Additional, rare, modes of transmission include the inoculation of contaminated biological samples, as urine, faces, plasma, or serum (79–82). Evidence shows that humans are able to infect other mammals, as cats, dogs or farmed minks, but just a few cases of inverse transmission have been documented so far (83–88).

SARS-CoV-2 can survive in the environment (e.g. stainless steel, plastic) from hours to a total of several days, depending on the type of surfaces (89,90). This period is increased by environmental conditions as low temperatures and low humidity, independently of surface type, and is reduced by sunlight (89).

Despite study heterogeneity, the proportion of asymptomatic infection among COVID-19 positive persons appears high (20-75%, in general population), supporting their epidemiological importance (91). SARS-CoV-2 incubation period ranges from 1 to 14 days, with a median of 5-6 days, which means an infected person may last 14 days from inoculation to the onset of symptoms (92). Nevertheless, the virus can only be transmitted to other individuals when replication-competent (i.e., infectious period). The infectious period begins 1-3 days before the onset of symptoms. The moment of the viral peak and highest infectiousness occurs in the first week (73). One modelling study, in which the mean serial interval was 5.8 days, estimated that infectiousness peaked between two days before and one day after symptom onset (93). The serial interval refers to the time between the onset of disease of successive generations of infected persons and can inform on the onset of infectiousness (37). Five to ten days after the infection, patients start producing neutralising antibodies, that bind to the virus and reduce transmission, but this process seems to take longer in severely ill or immunosuppressed individuals (73). Multiple studies have shown that the infectious period, based on positive cultures, was variable but rarely longer than 9 days after the symptom onset in mild cases and 20 days for hospitalised patients (73). The median duration of viral shedding was found to be of 18 days in a systematic review (94) but it can vary and may increase with age and the severity of illness (95–101). Also, persistent SARS-CoV-2 infection (marked by intermittent positivity) in immunocompromised patient can last several months (102–104).

The risk of SARS-CoV-2 transmission varies according to the contact patterns in the

population, depending, for example, on the type and duration of exposure, use of preventive measures (stay at home orders), and likely individual factors (viral load) (105–107). In fact, a minority of primary cases results in a majority of secondary infections, with previous studies showing that 80% of secondary infections traced back to 5-19% of primary cases (105–107). Certain settings increase the closeness and duration of contacts because of prolonged contact in indoor environments, as for example household, where the secondary attack rate was estimated to be around 18.1% (95% CI: 15.7%, 20.6%) or healthcare settings, where the secondary attack rate was 0.7% (95% CI: 0.4%, 1.0%) (108). Work settings are commonly associated with outbreaks as well, since they increase contact and sometimes include high-risk situations as meetings and sharing meals (109). Additionally symptomatic index cases present higher secondary attack rates than asymptomatic cases (Relative Risk: 3.23; 95% CI: 1.46, 7.14) (108).

Diseases that provide long-term immunity tend to reduce their transmission with time, since the susceptible population decreases. In the case of COVID-19, the disease seems to provide some kind of immunity but it is not life-long: some studies found previous infection to reduce the risk of re-infection in the subsequent six to seven months by 80-85% (110,111). For the variants of concern (e.g. B.1.1.7.) the risk of reinfection with is uncertain (112,113).

iv. Diagnosis

Early in the pandemic the viral whole-genome sequence (strain Wuhan-Hu-1) was characterised and released (114), allowing the development of Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) diagnostics test, the study of the virus replication, pharmacological testing, studying the pathogenesis and understand mechanisms that were the basis for vaccine development (2,115). Since clinical presentation of this disease is nonspecific, laboratory testing is of utmost importance in identifying cases and subsequently implement measures aimed at reducing viral transmission. In Portugal, two laboratory tests are accepted to confirm the diagnosis of present disease SARS-CoV-2 nucleic acid or antigen detection, respectively, PCR, or rapid test) (116). RT-PCR is the most widely used method for the diagnosis, because of its higher sensitivity, making it the gold-standard (117). However, the rapid test also presents advantages since it is cheaper and provides quick results.

v. Management

Health sector response to COVID-19 followed the same framework as other pandemics.

It should start with prevention, when epizootics is supposed to be controlled avoiding the spill over to the human species, followed by containment, with the rapid characterization and isolation of first cases to a small geographical area, and finally mitigation, with strategies to minimize the impact of the pandemic. Response strategies include epidemiological surveillance and risk evaluation, public health measures, delivery of healthcare services as inpatient and outpatient care, vaccines and other drugs, communication and evaluation (118). The methodologies used for the mentioned strategies can be divided into pharmacologic or non-pharmacological interventions.

Pharmacological interventions

The disease management is mostly symptomatic as there is no clinical evidence of effective antiviral drugs, and so no specific treatment (119). Several pharmaceuticals have been studied or are still undergoing clinical trials to assess their safety and efficacy as potential treatments for COVID-19, including corticosteroids, the antiviral nucleotide analogue remdesivir, systemic interferons and in particular interferon β-1a, monoclonal antibodies against components of the immune system such as Interleukin-6 (IL-6) and IL-4, other immune modulators and antibodies against components of SARS-CoV-2 (119). Lopinavir/ritonavir and Arbidol also bring little benefit (120). Tocilizumab, a monoclonal antibody, did not show statistical significance for all-cause mortality of severe COVID-19 patients when compared to the control group, with similar results for other outcomes (121). Remdesivir is the most promising antiviral drug for severe COVID-19 patients, showing a reduction in the 14-day mortality rate (RR=0.64) (120). Additionally, a meta-analysis has shown an association between systemic corticosteroids with lower 28-day all-cause mortality, when compared with usual care or placebo, among which dexamethasone had the strongest effect (122). Finally, Molnupiravir, an oral antiviral, has demonstrated a significant benefit in reducing hospitalization or death in mild COVID-19 (123).

Vaccines are a key tool in public health (119). COVID-19 vaccines aim to prevent COVID-19 infection, hospitalisation, severe disease or death, by triggering an immune response. Thanks to an unprecedented effort, information sharing, and bureaucracy reduction, it was possible to produce several vaccines against COVID-19 in a record timespan. As of August 2021, Portugal has 4 available vaccines: SPIKEVAX® (124), VAXZEVRIA® (125), Janssen® (126), and COMIRNATY® (127). The effectiveness showed to be high in real-live situations, with COMIRNATY® preventing 97.0% of symptomatic disease, severe/critical disease and death and 94.0% against asymptomatic SARS-CoV-2 infections (128). The MODERNA® vaccine showed 94.1%

efficacy at preventing Covid-19 illness (129). As for AstraZeneca®, the overall vaccine efficacy more than 14 days after the second dose has shown to be 66.7% (95% confidence interval: 57.4% to 74.0%) (130). A meta-analysis that assessed the global vaccine effectiveness, for all the four previously mentioned vaccines, found that they prevented any infection in 66.9% (95% CI: 58.4–73.6) (131). When the outcome was symptomatic infection, the pooled vaccine effectiveness was 75.7% (95% CI: 69.3–80.8), as for prevention of severe disease and hospitalisation was 93.8% (95% CI: 83–98) (131). Although preliminary results on vaccine effectiveness seem promising, doubts still remain on two main aspects: a) Are these vaccines effective to reduce COVID-19 transmission? Investigations are ongoing to assess further the impact of COVID-19 vaccination on transmission (132), but some studies already found reduced likelihood of transmission to household contacts (133,134). Additionally, a decrease in the viral load, deemed as a proxy for reduced transmissibility, has been shown for the vaccines against COVID-19. Yet, such reduction is not a perfect measure of reduced transmissibility (135–139); b) Are these vaccines effective against new variants? They are expected to provide at least some protection against new virus variants as they elicit a broad immune response involving a range of antibodies and cells (140). Therefore, changes or mutations in the virus should not make vaccines completely ineffective. In the event that any of these vaccines prove to be less effective against one or more variants, it will be possible to change the composition of the vaccines to protect against these variants (140). For example, evidence from July 2021 suggests the mRNA COVID-19 vaccines (Pfizer-BioNTech and Moderna) protect against most strains, including B.1.1.7 (Alpha, originally identified in the United Kingdom) and B.1.351 (Beta, originally identified in South Africa), however with reduced protection (141). In what comes to the B.1.617.2 (Delta) variant, evidence from October 2021 shows that compared with the Alpha variant, vaccine effectiveness against mild outcomes was reduced by 10-20%, but fully maintained against severe COVID-19 (131).

Non-Pharmacological interventions

In the absence of effective treatment options or widely implemented vaccination, non-pharmaceutical outbreak control measures aim to reduce transmission and therefore incidence. Two case-level public health measures are paramount to reduce transmission (as furtherly explained in Epidemiological investigation and contact tracing) (142): identification and isolation of infected cases, followed by tracing and restriction of movements of their contacts. By immediately isolating infected cases, further contacts are being minimized and therefore secondary cases are prevented. Tracking, tracing and

isolating contacts of COVID-19 cases is an effective public health measure for the control of the disease (143). In an ideal scenario, during the epidemiological investigation, public health personnel promptly identify people who had been in close contact with a primary case, and then contacts them to evaluate and isolate whenever appropriate. By doing so they can rapidly identify and isolate secondary cases that may arise after transmission from the primary cases, and therefore interrupt further transmission.

Other non-pharmaceutical control strategies include individual, organisational, and population-level measures. On the individual and organisational levels, reinforcement of infection prevention and control measures are implemented. Among those measures, hand hygiene, use of personal protective equipment (e.g. masks), respiratory etiquette, environmental cleaning and disinfection, air renovation, maintaining physical distances and avoidance of close, unprotected contact with people with fever or respiratory symptoms are included (144). Specifically on organisational level, some policies can facilitate case isolation, as “sick-leave” in which workers have the right of paid sick days and therefore are less reluctant on staying home when symptomatic. On the country level, the interventions to control the outbreak include (145–147): i) measures that support the ones in individual and organisational level, including ‘stay-at-home’ recommendations for risk groups or vulnerable populations, ‘stay-at-home’ recommendations for the general population, use of protective masks in public spaces/ public transport, teleworking recommendations and closure of workplaces; ii) more restrictive measures, namely stay at home orders, lockdown, travelling restrictions, mass gathering cancellations, closure of public spaces and closure of educational institutions. Other country level measures can increase the compliance of quarantine among exposed individuals, as the payment of lost working days by a social security fund.

In the first year of the pandemic, available drugs were not effective in reducing transmission and their role in reducing mortality and sequels was still unsatisfactory. Also, during the first wave of COVID-19 the vaccination was not available to allow herd immunity and considerably reduce transmission. These conditions enhanced the focus on epidemiological investigation and public health measures to control the outbreak. Since this study is focused on epidemiological surveillance, this topic will be presented in more detail.

c) Epidemiological Surveillance

Surveillance is the continuous and ongoing collection, management, analysis, interpretation, and communication of data (31,37). This is a necessary process of epidemiologic practice essential to obtain reliable information on the status of diseases,

their determinants, or health status of a certain population. From this information, public health professionals are then able to design strategies to promote health, prevent or control disease, and following implementation, evaluate them.

Epidemiologic surveillance is used in all three areas of disease, communicable, non-communicable and injuries, but also in the field of the health determinants that influence the previous, namely, socioeconomical, environmental, occupational, or behavioural determinants. The most common objective is to monitor the occurrence of disease and identify health events, but it can also be used to estimate incidence and prevalence rates, if the sample is representative (37). Surveillance is able to identify a phenomenon and indicate which groups are most affected, thus enabling a more precise action. The target population of a surveillance system is defined according to the information needs. For public health officials, the target population is usually the residents within their geographical jurisdiction, or people that are staying there (for example, for vaccination surveillance).

Surveillance systems are networks of people and activities that maintain this process with practical, uniform, and rapid methods (31,37). Data sources used in surveillance systems can include death certificates, hospital records, primary healthcare records, administrative records from healthcare services, sentinels, notifications, laboratory diagnoses, vaccination data, social security data on labour absence, information on vectors' dispersion, biobanks, online search terms, among others (31). It may be established at all geographical levels, from local to international, and the main goal is to observe or anticipate changes in trends to trigger the appropriate action, which usually includes investigation and control measures.

Surveillance systems allow the monitoring of health events and can rapidly identify public health threats, leading the implementation of control measures. For a surveillance system to be viable on long-term, a balance between the necessary information and the limits of feasibility in data collection is required (37). On the one hand, passive surveillance uses available data on disease notification requested by the government or local health authority and filled by physicians, who are held accountable by the information and have the responsibility of reporting (148). Due to its characteristics these systems generally lack completeness and quality, which can be minimised by simplifying the notification and briefing it. In Portugal, the National Epidemiological Surveillance System (SINAVE) (see Portuguese epidemiological surveillance system) is the surveillance system for infectious diseases and other public health threats. It is a passive system in which physicians at all levels (primary care, hospital, others) and sectors

(public or private) have the responsibility and obligatoriness of notifying certain diseases, including, for example, HIV, viral hepatitis, listeriosis, dengue, legionnaires disease, tuberculosis, measles, antimicrobial resistance microorganisms, and many others (149). This part of the system, which consists in the medical notification, is referred to in the tool as “SINAVE MED”. Disease notification is complemented with laboratory-based surveillance, which is referred to “SINAVE LAB”. Using laboratory diagnosis is highly effective for some diseases and carries a relatively low additional effort for personnel if a informatic system is in place (37). For example, in the surge of a new public health threat, the use of previously implemented or the rapid implementation of surveillance systems is fundamental to identify transmission chains and have an effective and rapid response, as in the case of COVID-19. Passive surveillance systems are relatively inexpensive and easy to implement. On the other hand, active surveillance comprises a system where specifically recruited personnel is responsible to identify new cases of a disease that have occurred (case finding) (148). They can do so by, for example, visiting healthcare facilities and directly see patients, interview physicians, review records, and surveying communities where a case index has been reported. Active reporting is generally more accurate than passive, but it brings higher costs (148).

Routine infectious disease epidemiology relies on reports of notifiable diseases, that rely on a set of uniform diagnostic criteria to allow comparison across regions or countries - case definition (31,32,148). The surveillance case definition is used to define a disease or status and is usually divided in three different areas: clinical, laboratory and epidemiological. It is intended to aid public health officials in recording and reporting cases and should be distinguished from the clinical definition. The latter is used by clinicians to diagnose individual patients requiring an initiation of their disease management, either to guide auxiliary diagnostic tests or treatment (150). These criteria can be clinical, epidemiological, or laboratorial (36). These reports then allow to obtain further information on the disease as the attack rate (proportion of the susceptible exposed individuals to an infectious agent who become clinically ill), the index case (the first case discovered by the health care system in an outbreak), and the primary or secondary cases, being the primary the case that brings the disease into a population and the secondary the individuals that where infected by it and that are able to infect tertiary cases, causing generations of infection (32).

For COVID-19, the case is defined based on epidemiological history (including cluster transmission), clinical manifestations (fever and respiratory symptoms), and results of SARS-CoV-2 nucleic acid or antigen detection. The case definition is not static and

changed several times since the beginning of the epidemics. For surveillance purposes, Portuguese Directorate-General of Health (DGS) considered a suspect case any person that developed (a) cough (new or worsening of the usual cough), or (b) fever (temperature of 38.0º Celsius or higher), or (c) dyspnoea, until the 9th of November 2020. Then, sudden anosmia and sudden dysgeusia were added (116,151). While the surveillance case definition is not intended to be used by healthcare providers for making a clinical diagnosis, in the case of Portugal, during the first months of the pandemic, a special telephone line (*Linha de Apoio ao Médico – LAM)* used the clinical criteria to determine access to testing (152). This compromised the diagnosis of asymptomatic, mild, or unusual clinical presentation cases and therefore biased the case detection.

i. International Health Regulations

In a globalised world, multi-country outbreak management is useless if not internationally coordinated. International Health Regulations (IHR) provide a legal framework that defines international cooperation standards in health security issues (153). It defines countries' rights and obligations in handling public health events and emergencies, and is overseen by the World Health Organisation (153). Last reviewed in 2005, the IHR is an instrument that legally binds 196 countries all over the planet. It establishes the implementation of standard requirements for points-of-entry, international travel and transport, and surveillance systems. Therefore it defines the core capacities that allow the detection, assessment, report and response to emergencies including: laboratories, human resources, surveillance, preparedness, response, risk communication, coordination and national focal point, national legislation, policy and financing (153).

The main motivation for the adoption of the IHR was the SARS-CoV-1 outbreak in 2002-2003, which started as an epizootic spillover in China and spread over 30 countries (154). The outbreak was relatively well contained, regarding the overall number of cases exported from China to other regions. However, it created an alert state worldwide and made it obvious the risk of globalisation and international travelling on the spreading of infectious diseases. The necessity for a global coalition to provide health security based on international standards was present and the IHR was reviewed. It was very similar to the Portuguese health authority's legislation, created in 1990 as part of the Basis Law of Health (155).

ii. Portuguese epidemiological surveillance system

In Portugal, health authorities are, usually, Public Health doctors that have the opportune and discretionary power of the state to protect populations' health in case of situations

that represent a serious risk to public health (155). They are hierarchically dependent on the health ministry and the directorate-general of health, and have to be publicly designated as such figures in the republic diary, their power is defined geographically, for which the position of the directorate-general of health includes the whole country (155). They are allowed to use all means necessary, in a proportionate way, to limit the risks and have competences in the area of epidemiological surveillance and the implementation of the IHR (155).

As mentioned in the “Epidemiological Surveillance” section, the Portuguese National Epidemiological Surveillance System (SINAVE) has the aim of identifying public health events and gathering, analysing, and communicating systematic data on public health risks (156). It also prepares contingency plans for situations of emergency. The system is supported by an electronic platform where physicians notify notifiable communicable diseases, local public health authorities fill the epidemiological investigation, and regional and national public health authorities review the data and supervise it. Recently, this system has also integrated laboratories, allowing the direct association between the clinical notification and the laboratorial one. Although compulsory, under notification has been shown in this system with high rates of incompleteness (157).

SINAVE allows the public health authorities to receive notifications promptly and to act fast. When receiving a notification, they perform the epidemiological investigation (which includes the epidemiological investigation and, for some diseases, the environmental one) and classify the case accordingly. The system does not allow a follow-up of the individual case or contacts since it is not designed to do so and moves to higher levels of the public health authority’s hierarchy when a case is classified. The system is theoretically built in a way that clusters can be identified and informed to other health authorities, but in practice the sharing of epidemic information in outbreak control is limited.

As explained, one important aspect in the epidemiological investigation is to identify the source of infection. In the case of COVID-19, the clinical notification asked about “contact with known or suspected case”. These could be contacts and not the source of infection. Therefore, the identification of the epidemiological link might not be precise with the current system.

iii. Outbreak investigation and management

When an outbreak is identified by the implemented surveillance systems it is necessary to identify the agent if it is not previously known (emerging infectious diseases), or the

source of an already known disease. Outbreak investigation should also identify risk factors for progression-related and transmission-related states for the infectious agents, which are useful to subsequentially guide the implementation of public health measures and patient management. One must be aware that outbreaks are unusual, and localized in time and space, limiting information collected. In particular, risk factors information collected during an outbreak may be biased, since cases outbreak-associated cases are more likely to be diagnosed and notified in opposition to sporadic cases (outbreak bias) (37). The ultimate of goal of outbreak investigation is its control.

The investigation and management of an outbreak/epidemics usually follow a systematic approach with the steps in Figure 4 (35).

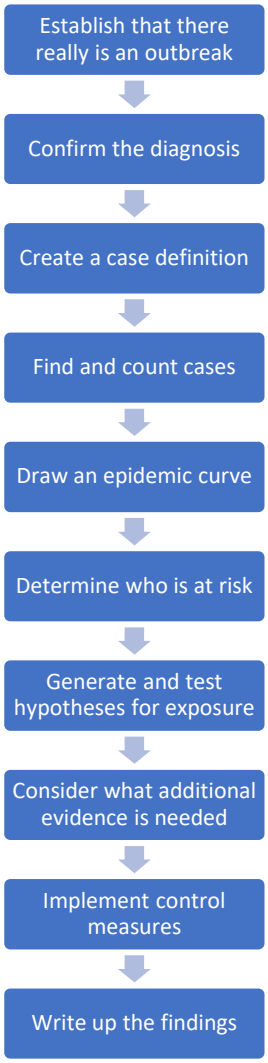

Figure 4 Steps of the investigation and management of an outbreak/epidemics. Source: Guest et al. Oxford Handbook of Public Health Practice - Oxford Medicine. 3 edn. Oxford University Press; 2013..

Epidemiological investigation and contact tracing

Local outbreak response is highly dependent on the Epidemiologic Investigation (EI) as part of the process shown previously. The purpose is to gather information on the specific case in a rapid and systematic way, which is essential not only to manage the case and contain the disease, but also to allow epidemiological surveillance and inform on broader prevention and control measures. Epidemiological investigation aims to identify causes and risk factors, implement prevention and control measures, and communicate with everyone involved (158). EI should also focus on measures of frequency of infectiousness and transmission. The definition of the degree and duration of the infectious state is therefore absolutely essential during the epidemiologic investigation of emerging diseases (37). Having a clear identification of the transmission chains is also essential to determine infectiousness. Other indicators of infectiousness include the secondary attack rate (the number of susceptible contacts of an infectious host that contracted the disease) (37).

The compulsory notification of certain infectious diseases is nowadays a public health practice consolidated in almost every country with comprehensive information systems (53). The procedures are usually defined by law and foresee that every physician that during his/her practice has knowledge of a suspected or confirmed case of one of the listed diseases or pathogens immediately communicates it to the public health authorities (53). The systematic collection of information is usually based on a formulary that is specific for each disease, according to its characteristics of clinical presentation, transmission, source, or others (53). When public health authorities receive the information from the physician, they first evaluate the case and define it according to specific clinical, epidemiological and laboratorial criteria in one of four possible ways: confirmed, probable, possible or not confirmed (53). It may use only the physician notification or may need extra sources as clinical, laboratorial, or environmental. When further individual information is needed, public health professionals can contact the physician and/or the patient or their families by direct interview, telephone, e-mail, post, or other available means (159). During the interview, public health professionals gather important information, such as, if there has been a contact with other known cases, the source of infection, when did the symptoms start, when was the person diagnosed, when did the person isolate from others, where the person has been, what the person has done (53). This allows public health authorities to understand if it is a sporadic case or if it is part of a cluster, and what is the source or the transmission chain. The investigation does not end in the interview, and complementary procedures may be necessary to

establish the source of infection, contacts, or important determinants, as for example further laboratorial studies of the patients' specimens. In the case of environmental source diseases, an environmental investigation is sometimes needed. It may include a detailed description of the physical places where the patient has been with the identification of critical points, complemented with laboratory tests of samples (for example, water samples in the case of legionnaires disease). If the environmental investigation shows important anomalies that could seriously influence the populations' health, immediate measures should be implemented (e.g., closure of an establishment). For some diseases, genotype analysis of the samples, both environmental and the patients' specimens, can prove essential on establishing a causal pathway and can be particularly useful in case of imported (or exported) cases.

In the case of person-to-person communicable diseases it is important to determine the infector with a minimum certainty level to implement measures and interrupt an undiscovered chain of transmission (160). The moment of contagion has to be in the window of the incubation period of the disease, thus, a detailed characterization of the patients' activities, contacts, or places frequented during the incubation period should be performed. The epidemiological link (source of infection) is identified by the public health professionals based on the interpretation of the data gathered, together with previous knowledge. It is not a precise attribution and may be indicated based on certain leads. It may be very clear, e.g., a family member previously diagnosed with the disease, or less precise, e.g., an association with a certain cluster, for example, a school outbreak.

Contact tracing is the process by which public health professionals identify individuals who have been exposed to a person infected (index case) with a pathogen that can be transmitted person-to-person(161). During the epidemiological investigation, and using the infectiousness period of time as reference, the infected patient close contacts are listed. After an in-depth characterization of the contact, and depending on the specific disease, public health officials stratify the risk, evaluate their clinical status, and implement necessary measures. These measures are dependent on the diseases' characteristics and can include isolation, prophylactic antimicrobial treatment, and vaccination. Diagnostic tests support the implementation of measures by determining contacts infection state (for example Tuberculin Skin Test). By identifying and intervening in the contacts, it becomes possible to interrupt the chains of transmission and control the disease. These strategies are in practice for many known diseases as tuberculosis, meningitis, smallpox, or COVID-19 (40). While this seems an attractive strategy, there are limitations. For tuberculosis, previous studies showed lower positivity rates among

identified contacts despite the large amount of resources put into contact tracing (162,163). For instance, tracing is highly dependent on the memory of the enquired and the willingness of the contacts to comply with the implemented measures. This latest limitation can be reduced by the public health authorities' power set in some countries like Portugal and though international coordinated efforts. Additionally, time restraints and the beginning of infectiousness before the onset of symptoms limit the effectiveness of contact tracing.

iv. Epidemiological surveillance in Baixo Vouga during the pandemic

The Public Health Units (PHU) in Portugal are operative units with responsibility in epidemiological surveillance (164). They are dependent on a double hierarchy: i) ministerial one, which is intermediated by Regional Health Administrations and Primary Care Clusters or Local Health Units; and ii) the health authorities, which is played by the Directorate-general of health and the Regional Health Authorities (164–166). These units are composed by public health doctors, who may or may not be health authorities, community nurses, health technicians, and other professionals working in the public health sector (as sanitarian engineers). In Baixo Vouga, the PHU has been systematically working with the national surveillance system SINAVE. SINAVE, as explained before, is the general surveillance system for notifiable diseases and depended on the clinical or laboratorial notification. In the early moments of the COVID-19 pandemics the system was not prepared for the laboratorial notification and the local health authorities often received lab results in an unofficial way. Additionally, the notification form for physicians was based on a model for other coronavirus, like SARS and MERS, which made it incomplete and inadequate for the SARS-CoV-2. The system, as explained earlier, is designed to allow the prompt classification of the case, and inform higher levels, and it does not allow a follow-up of the case or its contacts.

When the COVID-19 pandemic arrived, the local area of Baixo Vouga was hardly stricken, which raised the need of an internal system to keep track of all the cases and its contacts. By then (March 2020), the national system (SINAVE) did not allow a recording of the contacts nor the visualization of intelligible information (e.g., incidence, Re, case-fatality, and others). It also did not allow the systematic laboratory reporting of pathogen genotyping data, namely genome sequencing data. Due to the high volume of notifications, SINAVE informatics platform was also experiencing technical issues. To bypass limitations of the existing system, the local Public Health Unit created a parallel system only used by its professionals. Later, at the end of March 2020, a new national platform was created to allow the management of COVID-19 confirmed cases, contact

tracing, and the vigilance of isolated patients in self-care – TraceCOVID-19 (167). This system could be accessed by all physicians and public health professionals, but it did not allow the PHU to access to local data. Additionally, the internal database contained data that was neither available in SINAVE nor in TraceCOVID-19, such as the epidemiological link (that was only added to TraceCOVID-19 in June 2020). The information obtained by the internal form was aggregated in a single database with information that was essential to manage the outbreak. The workflow in the PHU of BVPCC, including the use of the three different platforms is shown in Figure 5.

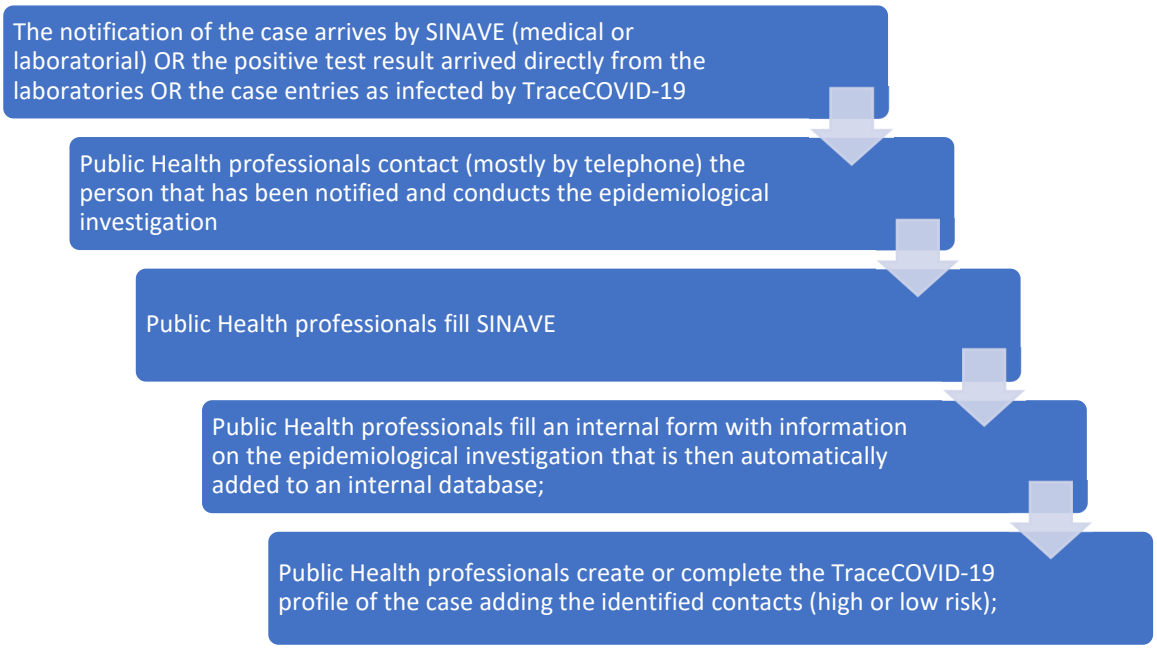

Figure 5 Working flow for epidemiologic investigation in the PHU of BVPCC.

d) Public Health Genomics

The use of genomic tools to inform public health is denominated Public Health Genomics. This field can be described as the "responsible and effective translation of genome-based knowledge and technologies for the benefit of population health" (Bellagio Group 2005; Burke et al. 2006). This is an emerging field, closely linked with various topics as precision medicine, precision public health, and pathogen genomics in infectious disease control and prevention.

In the last two decades great steps have been made in the human genomics field, with the release, in 2003, of the "reference sequence" for the human genome, including the almost complete sequences of around 25,000 genes (168). These advances also provided important tools for the research of gene-disease associations and contributed to the understanding on human genomic variation and its interactions with the

environment to improve health and, prevent and manage disease. Genomics can be divided into three areas, at two levels of intervention. Two areas occur at an individual level, diagnostics, and treatment, and one at a population level, risk-stratification.

On the one hand, in the individual patient management area, genomics provided a shift to a more precise medicine. Typically, medicine uses “average patient”-oriented approaches, based on epidemiological research to establish the mean population characteristics, and then applies interventions to a certain individual which is very likely not to be exactly in that mean. Although still lacking evidence, precision medicine could provide diagnostic and treatment tools specifically tailored for a real individual (169–171). For example the implementation of pharmacogenomics is notorious, since it allows a personalized treatment informed by pharmacogenetic testing (170,171). While these are predominantly being applied in the cancer treatment area there is room for expansion.

On the other hand, to promote health and prevent disease at a population level, genomic information offers the opportunity for a proactive action, contributing to precision public health (172–175). At this level, genomics helps identify pools of individuals with certain genotypic characteristics that increase the risk of disease, that can then be an intervention target to modify environmental factors, such as diet (40). Genomics can provide more accurate methods for measuring disease, pathogens, exposures, behaviours and susceptibility, which can lead to a better evaluation of the population’s health and a better choice of the necessary interventions. As in the case of precision medicine, public health genomics drove the development of public health precision, defined as providing the right intervention, to the right group of people, at the right time (101). There are limitations to these interventions as, for example, the fact that the penetrance frequency of certain predisposing genotypes is not known, making it hard to understand the real risk associated to the development of the disease when the genotype is present. In this case, epidemiology can help by giving broader information of prevalence and population screening. Additionally, some controversy exists on the ethical issues regarding precision public health interventions, since it would mean to focus more efforts and resources in a certain group of individuals to achieve a greater gain, instead of focusing in all the population achieving less individual gains but a greater one when summed up (41). Another concern is that the increased emphasis on genomics may take the focus out of changing other health determinants (social, environmental, and others) (174). In fact, some determinants of health are very hard to change (behavioural, social) and public health has been struggling to fight them, mostly without success. But putting the focus only on genomics could also risk being too deterministic in the health

promotion approach, putting the blame only on the genes and taking the responsibility out of the individual itself. In short, the field of precision public health uses large-scale data and technologies, including genomic, spatial analysis and big data, to better identify determinants of population health that can be targeted through precise interventions and ultimately lead to an improvement in populations' health outcomes (177). Despite presenting some unsolved ethical issues, it appears as a promising, innovative field.

i. Pathogen Genomics

The surveillance process is not limited to the detection of cases of a disease and characterization of epidemiological characteristics. Sequencing information of the pathogens can aid epidemiological investigation by identifying epidemiological links more accurately. While public health precision and the use of genomic tools are still under definition, it is clear that genomic epidemiology is able of identify either population dynamics or individual transmission events (178,179). Since the technology is becoming cheaper and simpler, it is possible to perform sequencing technics in field situations with low resources. These tools have been implemented through outbreaks settings, community settings, acute care hospitals, and other healthcare facilities (180). Biosurveillance can detect new pathogens that cause disease in humans. With the globalisation, climate changes and the continuous human-animal interactions, the potential spread of already existing diseases and emergence of new ones increases. To manage these infectious agents, appropriate knowledge is necessary on its clinical and epidemiological features, and transmission characteristics.

The field with a well-established application in medical and public health is the one of Pathogen Genomics (181). It integrates sequencing along with bioinformatics to determine the distribution and spread of a communicable disease in a certain population and the use of this information to implement control outbreak measures (182). By genotyping a pathogen, it is possible to ascertain it to a phylogenetic level, based on their genetic relatedness. Whole-Genome Sequencing (WGS) is a widely used method in surveillance for is replicability, allowing integration of data from different contexts (e.g., countries).

This field is already in place in many countries, including Portugal and continues to expand (183). It uses molecular technologies to provide guidance to the diagnosis, treatment, and control of communicable diseases. For example, metagenomic sequencing of respiratory samples from first COVID-19 patients in China allowed the identification of SARS-CoV-2 as the causative agent of severe pneumonia cases (114), which was pivotal not only to develop specific diagnostic tests, but also to implement

global genomic surveillance networks able to track the virus evolution, adaptation and dissemination(184,185). These initiatives combine genome technologies with bioinformatics and epidemiology to enhance and accelerate public health surveillance, investigations, and control of infectious diseases (170,182). (49,160). Pathogen genomics is a useful tool in several areas from food safety, animal health, environmental health, and human health. The integration of multiple sectors is essential to effectively control and manage infectious diseases, recognizing “One Health” approach. For this reason, international agencies, such as the World Health Organization (WHO) and the European Center for Disease Prevention and Control (ECDC), strongly promote the integration of genome sequencing data, from multiple sectors, as an essential component of surveillance systems. Multiple pathogens are nowadays monitored and investigated through genome sequencing, such as influenza (182,183,186), *Escherichia coli* (187,187–189), *Legionella pneumophila* (190), *Salmonella* (191–193), *Listeria monocytogenes* (194), *Staphylococcus aureus* (195), *Mycobacterium tuberculosis* (196–200) and, recently, SARS-CoV-2 (9,10). In Portugal, some uses that are already in place include respiratory infections (183,201–203), sexually-transmitted diseases (204,205) and foodborne pathogens (206).

ii. Pathogen Whole-Genome Sequencing in Outbreak Investigation

Whole-Genome Sequencing (WGS) allows the identification and characterization of pathogens and helps resolving disease transmission patterns during outbreak investigation. In particular, continuous WGS of virus can help mapping the transmission across communities, study mutations for their potential to cause increased transmission or virulence, alert public health authorities to superspreading events or hotspots, help identifying introductions of new variants and point to possible sources of infection (182). These inferences are based on the rationale that virus mutate while they spread, leaving a genomic fingerprint that can then be used to infer ancestral relationships (phylogeny) (207). When used for virus, this inference is based on the rationale that virus mutate while they spread, leaving a genomic fingerprint that can then be used to infer ancestral relationships (phylogeny) (182). WGS is currently performed using advanced and high-throughput technologies (so called, next-generation sequencing technologies - NGS) to detect these fingerprints in an inexpensive and in-depth way, which means it is possible to get sequences on a massive scale (182,208). WGS has also the potential to determine whether an outbreak is of natural causes, or is due to deliberate or accidental introduction of an engineered or known pathogen (208). Additionally, WGS also proved to be useful in providing accurate and rapid information to report data on urgent events, such as

healthcare associated outbreaks (182).

In brief, pathogen WGS in the context of routine surveillance and outbreak investigation comprises the following main steps (182,208):

1. Collection of patients sample;
2. Inactivation and extraction of genomic material, i.e., nucleic acids (RNA/DNA);
3. Preparation of nucleic acids for WGS through sequencing library preparation;
4. Sequencing of the latest library using a sequencing platform using either short-read or long-read technology);
5. Multi-step advanced bioinformatics analysis, involving i) quality analysis and improvement of raw sequencing data; ii) reconstruction of the pathogen genome sequence using reference-based mapping and/or de novo assembly approaches; iii) detection of specific genetic markers (e.g., determinants of antimicrobial resistance or virulence); and iv) phylogenetic inferences.

The obtained genomic data should then be combined with other epidemiological evidence to allow the reconstruction of transmission history, that can then be useful to decision and policy making (208). Genomic epidemiology is based on the assumption that it will not be possible to sequence all the cases of disease in the transmission network while still being able to give key information on the epidemic process (182). This is possible since connections between sequenced cases can be made, based on distance-matrixes that calculate the genetic distance from multiple sequence alignments, determining their phylogenetic relationships. By allowing the identification of strains that are more or less closely related to each other, this process is able to identify cases more likely to be epidemiologically linked, and this is fundamental to make a comprehensive analysis of the outbreak transmission dynamics.

A phylogenetic tree can be represented in different ways, such as rectangular rooted trees (more common), radial rooted trees, or unrooted trees. In this tree (Figure 6) several elements are identified and depicted: i) the leaf nodes, which are the sequenced samples; ii) the outgroup, which is a node that represents a sequenced sample that is less related to the others; iii) the horizontal branches of the tree that represent genetic distance (commonly expressed as Single Nucleotide Polymorphims - SNP); iv) the internal nodes, which are common ancestors (not sampled); v) the root, which is the common ancestor to all other nodes and vi) the clade, which is a group of closely related sequences that share an internal node (182). For SARS-CoV-2, distinct clades of the

global phylogeny, which are defined by distinctive signatures of “shared mutations”, have been conventionally named as clades (e.g., using the Nextstrain or GISAID clade nomenclatures) and Lineages (using the Pango Network nomenclature). The term “variants” has been applied to the SARS-CoV-2 has a synonym of clades or lineages of closely related virus defined by specific constellations of mutations (182). Spike protein mutations have been especially studied because of their importance on the host-receptor binding and vaccination targeting. In general, the global burden of variants changes in time and in place, with some variants raising their frequency until dominance at local or global levels, can be explained, for instance, by a founder effect or altered viral properties that give them transmission advantages (e.g., increased viral load, increased transmissibility) over other circulating variants. For instance, in the specific case of Aveiro region, the genomic analysis showed a wide spread of a particular variant (carrying the Spike mutation D839Y), which was most likely driven by a strong founder effect of a potential single introduction (182). Variant identification plays a small role in transmission chains identification but is clinically and epidemiologically important. In the specific case of Aveiro region, the genomic analysis of the samples showed a wide spread of a particular variant (D839Y), that supported a strong founder effect of a single unique introduction (209).

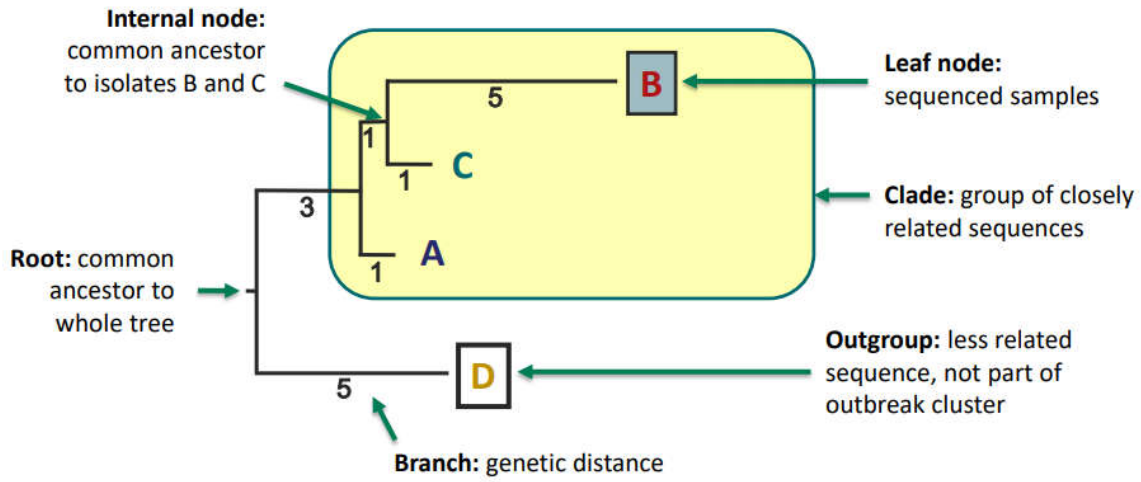

Figure 6 Anatomy of a phylogenetic tree. Source: Advanced Molecular Detection (AMD) and Response to Infectious Disease Outbreaks (20).

Noteworthy, although extremely useful in epidemiological investigation, phylogenetic trees need to be complemented with additional data and they cannot fully resolve transmission pathways as sequenced samples are only a fraction of the entire transmission chain, and the existing gaps can change its interpretation (182). Phylogenetic trees are thus an approximation of the truth, and its degree of accuracy is

associated with the samples sequenced.

With the pandemic of COVID-19, pathogen genomics has been applied to the SARS-CoV-2 virus surveillance throughout the world. By comparing sequences for a certain region to others around the world, it has been possible to better understand the introduction and transmission dynamics of the virus between and within countries, with benefits for pandemic control. Nevertheless, although SARS-CoV-2 genomics has shown to facilitate public health decision-making (22,210), there are still huge discrepancies between countries. In fact, comprehensive genomic surveillance systems involving systematic integration of epidemiological and genomic data are only completely established in a few places due to the multitude of resources needed (182,211). In fact, genomics-informed public heath outbreak responses for SARS-CoV-2 require a wide participation of laboratories, (de)centralized data analysis (depending on the geographical organization of the health system: national, regional or local), efficient information systems allowing a timely integration and sharing of genomic and epidemiological data, etc. In Portugal, such framework is partially established but still needs to be strengthened to facilitate the systematic integration of genomic and epidemiological data and rapid sharing of results back to local public health entities.

e) Use of genomic and epidemiological data in COVID-19 epidemiologic investigation (Genomic Epidemiology)

As explained before, Epidemiological investigation and contact tracing, is deemed essential to find transmission chains and control COVID-19. Pathogen genomics can help furtherly understand where the epidemiological investigation failed to identify accurately the transmission chain. Inversely, SARS-CoV-2 sequencing alone can be limited in providing the source of infection, its context and direction. When put together, these two disciplines have the capacity of overcoming those difficulties in most cases, providing detailed information (from genomics) and putting it into context (through epidemiology) (212–214). Current knowledge on the agreement between whole-genome sequencing and epidemiological data on the link of COVID-19 cases (Table 1) is scarce(22,23,25,28,215,216). Evidence on quantitative data is limited. Seemann et al. reported that that for each epidemiologically-linked group the (epiclusters identified by the epidemiologic investigation) a median of 100% (Interquartile Range (IQR): 93–100%) of cases associated with a single dominant viral genetic cluster (identified on the basis of viral genomic data) (22). Additionally, they estimated the agreement in the inverse direction and identified a median of 43% of cases (IQR 23–77%) in each genomic cluster to belong to a single dominant epidemiologically-linked group (22). An Austrian study

(Popa et. al) combining WGS and epidemiological data found an overlapping of 65% between these two types of data sources (215). In a healthcare setting in the United Kingdom, Snell et. al identified agreement in 75 out 168 (45%) of epidemiologically linked cases, for which sequenced cases were not refuted by applying genomic data (216). They described some factors associated with the epidemiological un-identification of transmission, such as a high proportion of community onset cases, clusters with cryptic transmission, and the delay on the identification of the primary case of the cluster (216). Other studies analysed from a qualitative point of view the coherence between epidemiologically and genetically defined clusters. Da Silva et. al were able to identify travel-associated introductions in Scotland by merging the two data sources, they also identified an international conference as an important outbreak source, which was effectively contained by health authorities (23). This study was not able to distinguish the direction of the infection using phylogeny, but they found that, for the majority of cases, the phylogenetic lineage placement correlated well with the available epidemiological introduction events (23). In Oman, a study combining epidemiological and whole-genome sequence data from 94 samples of SARS-CoV-2 was able to find a robust surveillance system, with a good agreement between the two data sources, that was very efficient in guiding the outbreak investigation processes in the country (25). Another study, from East England, analysed 299 cases in healthcare settings and identified 35 different clusters (28). From those cases, 58% had strong epidemiological links and 20% had plausible epidemiological links (28). This data was fed back to clinical, infection control, and hospital management teams, leading to infection-control interventions and informing patient safety reporting (28). Furtherly, other studies were important in confirming or excluding transmission in particular settings. Page et. al, identified a discrete sublineage associated with six care facilities and found no evidence of reinfection in longitudinal samples; ruled out a healthcare associated outbreak (217). A small amount of studies have been published describing the use of WGS to infer transmission chains in the Portuguese context (218–220). Other two studies have been carried in a country-level and province-level lockdowns, in Thailand (24) and Italy (26), respectively. This is a special context since its expected no transmission between the enclosed space and other geographical regions. Epidemiological investigation and control measures rely heavily on epidemiological enquiries. Thus, knowing to what degree and in what contexts epidemiological investigation is failing to identify the source of infection has important consequences on the subsequent implemented measures of contact tracing and isolation.

Table 1 Literature review on studies combining whole-genome sequencing and epidemiological data of COVID-19 cases.

Article	Use of WGS and epidemiologic data	Concordance analysis? (Yes/No; Sensitivity of epiclusters; Sensitivity of genomic clusters (%))	Genomics used to detect links	Sample size (Epidemiological/ WGS)	Framework for the application of genomics	Assessment of Public Health interventions	Study period	Geographical area in sanitary cordon	Country
Seemann et al.(22)	Yes	Yes; 100% (IQR 93–100%); 43% (IQR 23–77%)	Yes	1 333 / 903 (68%)	Yes	Yes	6 January - 14 April 2020	No	Australia
da Silva Filipe et al.(23)	Yes	Yes (qualitative); "for the majority of cases, the phylogenetic lineage placement correlated well with the available epidemiological introduction events"	Yes	2 641 / 1 314 (50%)	Yes	Yes	28 February - 31 March 2020	No	Scotland
Puenpa et al.(24)	Yes	No	No	(not available) / 40	No	Yes	12 January - 17 May 2020	Yes (country-level)	Thailand
Al-Mahruqi et al. (25)	Yes	Yes (qualitative); "Good agreement"	Yes	456 / 2171 (21%)	Yes	No	24 February - 23 May 2020	No	Oman
Giallonardo et al. (26)	No	No	Yes	(not available) / 191	Yes	Yes	January - July 2020	Yes (province-level)	Italy
Böhmer et al. (27)	Yes	No	No	16/16	No	No	20 January – 11 February 2020	No	Germany
Meredith et al. (28)	Yes	Yes (qualitative); A single large outbreak in the hospital presented incongruent lineages. A single lineage in a care facility outbreak. High (100%) congruence in a food processing facility.	Yes	374 / 262 (70%)	Yes	Yes	13 March – 24 April, 2020	No	England (hospital settings)
Popa et al. (215)	Yes	Yes; 65% sequenced cases assigned to epidemiological clusters	Yes	345 / 10 385 (33%)	Yes	Yes	24 February - 7 May 2020	No	Austria
Page et al. (217)	Yes	No	Yes	1305/1376	Yes	Yes	March - August 2020	No	East England
Snell et al.(216)	Yes	Yes; 45%	Yes	168/234	Yes	Yes	10th March - 31st April 2020	No	United Kingdom

Sanitary cordon refers to an area in complete lockdown, where access was allowed only for essential reasons – healthcare, provisions, police. Legend: IQR (Inter Quartile Range)

f) Pertinence, investigation question and aim

The COVID-19 pandemic has imposed a great stress in health systems and in the countries' economies. Until vaccination was completely implemented, the only intervention available to control it is to reduce transmission by stopping transmission chains. Knowledge on how transmission occurs is of extreme importance. There has been a rapid evolution of the disease and Portugal has struggled to control it. In the need to provide further information, this study presents a genomic epidemiological analysis describing COVID-19 transmission chains, from March to June 2020, in Aveiro Region, an area with a municipality in sanitary cordon, using epidemiological and genomic data. By revisiting a large epidemiological investigation (1925 COVID-19 cases), carried out in one of the regions with the highest incidence rates during the first epidemic wave in Portugal, through the integration of available viral genomic data. The magnitude of this epidemiological investigation, involving the only mainland region subjected to a cordon sanitaire, coupled with the use of detailed viral genomic data to dissect transmission networks is unprecedented in Portugal. Besides the assessment of the concordance between epidemiologically linked cases (epiclusters) and viral genome data, allowed to unveil hurdles and challenges behind the misidentification of epidemiological links, contributing to the knowledge on the transmission dynamics and the accuracy of the epidemiological investigation of COVID-19. Additionally, it consolidates the expected benefit of pathogen genomic surveillance, as well as the identification of new opportunities to improve the preparedness for future epidemics. Regardless of the retrospective character of the present study, it provides detailed investigation, and the study outcomes are very opportune and informative in the context of the current technological revolution of surveillance and information systems at global level. Also, the results can particularly contribute to changes in the frameworks of integration between field epidemiology and pathogen genomics, at national level, as well as in Public Health clinical practices, by raising awareness among public health actors and decision-makers about the need to embed genomics into the routine activities of infectious disease surveillance and outbreak management.

i. Research question

Are epidemiological and viral whole-genome sequencing data in agreement when used to describe the transmission chains of COVID-19 in Aveiro region, from March to June 2020?

ii. General Objective

Describe COVID-19 transmission chains, in Aveiro region, from March to June 2020, using

epidemiological and viralwhole-genome sequencing data, and describe its agreement.

iii. Specific objectives

1. Rebuilt COVID-19 transmission chains, in Aveiro region, from March to June 2020, using epidemiological data and whole-genome sequencing data.
    a. Rebuilt the transmission chains;
    b. Describe transmission dynamics and settings of the main epiclusters.
2. Describe the agreement of whole-genome sequencing data for each epidemiologically linked chain of transmission:
    a. Estimate the frequency of cases within and without the main viral genomic profile, for each epicluster;
    b. Estimate the sensitivity of the epidemiological investigation to detect links between cases that belong to the same transmission chain, for each epicluster;
    c. Estimate the pooled sensitivity of the epidemiological investigation to detect links between cases that belong to the same transmission chain.
3. Review sporadic cases and misidentified direction of transmission, according to available viral whole-genome sequencing data:
    a. Describe and characterize misidentification of epidemiological links in terms of direction, using viral WGS;
    b. Describe, characterize, and analyse cases identified as sporadics by epidemiological investigation, using viral WGS.

3. Methods

a) Design, setting, data sources, and data collection

i. Design and setting

This is an outbreak analysis study using previously collected data on epidemiological and genomic characteristics of all confirmed cases of COVID-19 prior to 30th June, in Aveiro region.

ii. Inclusion and exclusion criteria

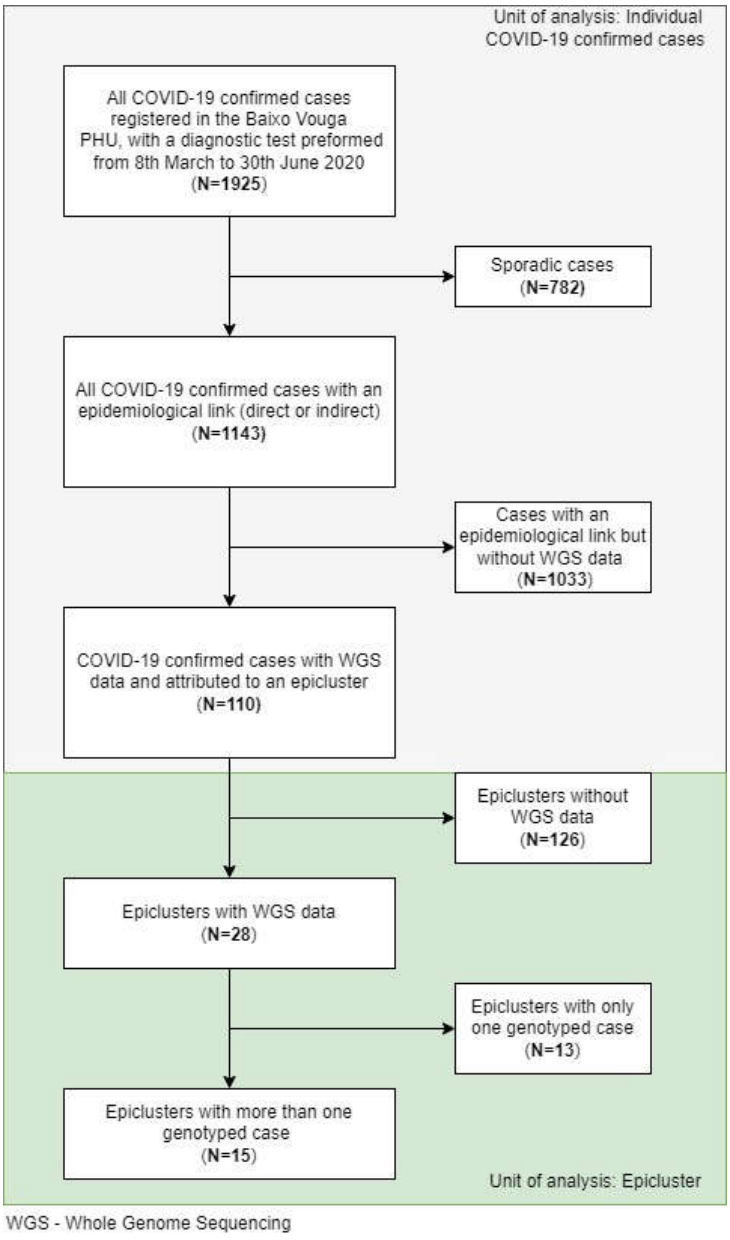

Figure 7 Flow diagram representing the sample selection methods and criteria of inclusion and exclusion

Inclusion and exclusion criteria are demonstrated in Figure 7, two units of analysis were used: individual confirmed cases of COVID-19 and epiclusters. All COVID-19 confirmed cases registered in the Baixo Vouga PHU, with a diagnostic test preformed from 8th March to 30th June 2020 (N=1926) were used. No duplicated cases or cases with missing information on the identification were found, due to previous database cleaning by the health authorities. To construct epiclusters only cases with an epidemiological link (i.e., with an identified infector, or cases with an identified subcluster) were included. When comparing epidemiologic and genomic data, to estimate the pooled sensitivity, only epiclusters with more than two sequenced cases were included.

iii. Epidemiological data

Data was collected from confirmed cases and respective epidemiological investigation by the PHU in Baixo Vouga. Confirmed cases were captured by the public health authorities through 3 possible ways: i) the medical or laboratorial notification through the national surveillance system platform; ii) the positive test result arrived directly from the laboratories; ii) COVID-19 tool for follow-up and contact tracking (TRACE-COVID-19). Each case had a unique Identification Number (ID). Data was collected during the epidemiological enquiry, by contacting the cases and respective contacts. Mainly, this data collection was performed by a phone contact and informatic databases, to obtain information on clinical presentation, clinical status, or test results.

Epidemiological investigation - variables

Information collected included detailed demographic (age, sex, and municipality), clinical (date of symptoms onset, clinical presentation, and medical conditions), laboratorial (data of the diagnostic test, type of diagnostic test), and epidemiological data (epidemiological link (infector and context), date of probable contagion, and designation (cluster or sporadic)). A detailed variables operationalisation plan can be consulted in Appendix 1.

Epidemiological links refer to the putative source of infection of a confirmed case using the epidemiological investigation, this duplet is hereby considered as the *infected/infector. Epicluster* refers to the complete transmission chain of interconnected cases, based on the infected/infector duplet. Epidemiologically linked cases were defined as (a) direct, when a confirmed case was identified as the probable infector OR when the infector was after identified through contact tracing (by testing the close contacts of the first confirmed case), and (b) indirect, when an infected confirmed case belonged to a context where a cluster was identified but there was no clear connection with one specific confirmed case as its infector. In the case of indirect epidemiological link, the infector was missing, so it was considered to be the (1) the primary case of the epidemiological cluster identified following investigation (the highest ascendent of the chain without an identified epidemiological link), (2) the first case of that epidemiological cluster to report symptoms, (3) the first case of that epidemiological cluster to present a positive test, (4) the first case of that epidemiological cluster to be identified by the PHU. The context of the epidemiological link was attributed accordingly to the relationship with the infector as cohabitant, healthcare, work settings, non-cohabitant (friends and other acquaintances), other, unknown (when the epidemiological link was not identified). The same epicluster could include smaller clusters (subclusters) identified by the health authorities during the epidemiological investigation. A

subcluster was considered in the presence of two or more cases at the same time and space.

Cases were considered sporadic in the following situations: i) epidemiological data was available but no infector was found; ii) epidemiological data was available but the epidemiological link involved cases outside the geographical area of the study; iii) there was insufficient or inexistent data referring to the epidemiological link. Sporadic cases were included in the descriptive analyses but excluded from the epicluster analysis (see below). For contagion context, individuals without an identification of either the infector or the epicluster were classified as unknown. Those with available information in the context of contagion were classified in cohabitant, healthcare, non-cohabitant acquaintances, work settings or other (individuals with identified context that could not be attributed to any of the above). Individuals identified with an epicluster or infector outside the geographical area of study were deemed as sporadic but the context was still assessed.

To analyse clinical presentation, individuals were considered symptomatic if either reported as symptomatic, having date of symptoms onset or presenting any symptom.

iv. Genomic sequencing and bioinformatics analysis

Sample collection and inclusion criteria

In Portugal, RT-PCR positive samples for SARS-CoV-2 from country-spread laboratories (enrolled in the Portuguese network) (209) are regularly sent to the National Institute of Health Doutor Ricardo Jorge (INSA) for genomic analysis, in the context of the genomic surveillance of the virus. The genome sequences analysed in this study were retrieved from the nationwide SARS-CoV-2 genome collection and, as such, no specific inclusion criteria were established to define the sample set. The availability of genomes associated with the studied patients was therefore contingent on the laboratory/Institute's partnerships in the genomics network, sample shipment conditions and sequencing success (strictly linked with viral load).

Baixo Vouga hospital was not a partner of INSA, and so, patients that only were tested in hospital settings were not covered for this study. INSA aggregated genomic data was provided as an intelligible database to the Baixo Vouga Public Health Local Unit. Since the data for this study was collected from the national SARS-CoV-2 genome collection, no specific inclusion, besides residing in Baixo Vouga geographical area was considered. No exclusion criteria were applied to retrieve the study sample. However, the availability of genomes associated with the studied patients was contingent on the laboratory/Institute's

partnerships in the genomics network, sample shipment conditions, and sequencing success (strictly linked with viral load).(221) Generally, high success rate in whole-genome sequencing is obtained for samples with a low cycle threshold (Ct). Ct refers to the number of cycles needed to amplify viral RNA to reach a detectable level, the lower the Ct the higher the viral load (82).

Viral whole-genome sequencing

SARS-CoV-2 positive RNA samples were subjected to amplicon-based whole-genome amplification with tiled, multiplexed primers, following the Artic Consortium protocol (222). Briefly, after Illumina NexteraXT library preparation, paired-end sequencing was performed either on Illumina MiSeq or NextSeq 550, targeting ~1M reads per sample. Analysis of sequence read data was conducted using the bioinformatics pipeline implemented in INSaFLU (183), which is a web-based (and also locally installable) platform for amplicon-based next-generation sequencing data analysis. Phylogenetic analysis and clade assignment was performed with Nextstrain (223). Lineage classification was performed using Phylogenetic Assignment of Named Global Outbreak LINeages (224).

Integration of viral genomic data with the epidemiological data

To allow the integration of epidemiological and genomic data, a team was created between the Public Health Unit of Baixo Vouga and the Bioinformatics Unit of INSA, including professionals from multiple backgrounds (public health physicians, epidemiologist, and microbiologists). The team held online meetings to enable interactive reporting of genomic epidemiological analyses and facilitate the translation of genomic findings into public health information.

To evaluate the phylogenetic context of the studied genome sequences and assess their genetic relatedness, phylogenetic trees were built for both the whole national genome collection and the Baixo Vouga dataset. To facilitate the congruence analysis between genomic and epidemiological data, genome sequences were grouped into “main viral genetic profiles” (i.e., “clade-level” clusters, 1-4) and further classified by alphanumeric Hierarchical Codes (HC) according to their ancestry within each main genetic profile.

b) Statistical analysis

i. Sample and setting description

Since sequencing was performed in a subset of the total confirmed COVID-19 cases, the

characteristics of the included sample in the analysis was compared to the individuals not sequenced, to assess selection bias. The sample was described using:

- Categorical data: absolute and relative frequencies, and Chi squared or Fisher test, depending on the expected counts per cell (if <5, Fisher test was used);
- Continuous variables:
    - mean, standard deviation and t-test or ANOVA, if normal distribution could be assumed following Shapiro-Wilk test for numerical data.
    - median, Interquartile Range (IQR) (i.e., 1st and 3rd quartiles), and Kruskall-Wallis, if normal distribution could not be assumed following Shapiro-Wilk test for numerical data (age).

ii. Epidemic curve

To describe the context, an epidemic curve was presented along with contextual details of all confirmed COVID-19 cases from March 8th to the end of June, in Baixo Vouga. The epidemic curve of Baixo Vouga was built using the number of daily new cases notified to public health authorities. Since, the variable symptoms onset contained missing data, a confirmed case was attributed to the date when the sample for diagnosis test was collected. The periods of meaningful public health measures that affected Baixo Vouga are represented as horizontal lines, including the sanitary cordon, emergency state and its announcement at a national level, country-level general lockdown and mass events restrictions.

iii. Transmission chain reconstruction

The denomination of *"Epicluster"* refers to the complete transmission chain of interconnected cases, derived from epidemiological investigation, where nodes are confirmed COVID-19 cases and edges are the epidemiological link between infected and infector duplets. The same epicluster may include many subclusters, identified during the epidemiological investigation.

The reconstruction of epiclusters was made using R (version 4.0.3) package Epicontacts (225). The connections between cases (edges) were based on previously generated duplets of infector-infected. Each node (confirmed case of COVID-19) had associated data on demographic, clinical, epidemiological and viral genomic sequencing characteristics. In the transmission chains provided by Epicontacts, nodes were coloured by subcluster and shaped by main viral genetic profile, and edges were coloured by the type of link (direct or

indirect).

If bidirectional links existed between two patients, the ascendant was considered as: (1) the individual that had the onset of symptoms first, or (2) the case that had an identified epidemiological link out of this group of two. Cases that did not report an infector (directly or indirectly) were excluded from the network analysis but are included in the descriptive analysis for comparison.

Aggregated data was then obtained using the epicontacts object in order to extract information on the characteristics of the epiclusters (i.e., identification of the epicluster, size of the epicluster, number of individuals in each main viral genetic profile).

The description of the transmission chains was performed in the two main epiclusters (A and B), corresponding to those with a higher number of attributed patients according to the epidemiological investigation. All single transmissions were verified according to the congruence between the hierarchical code of the infector and the infected. When WGS data was not available for both cases in the duplet close cases presenting such data were considered. For example, in a case of presumed transmission between a healthcare professional and a patient, if there was no information on WGS samples of the professionals, a familiar or colleague derived WGS data with a high probability of link was used, assuming the professional's one was similar. Most of the times a clear transmission chain from infector to infected was not possible to unravel, so subclusters to follow the chains were used. If a subcluster had cases with a viral genetic profile coherent with the epidemiologically linked ascendant, it would have been ascertained as confirmed in the transmission chain. Viral genetic profiles of genotyped cases were considered to describe the consistency of the epidemiological chain, as ascertained by public health authorities.

iv. Agreement analysis

To answer the research question, a method to estimate the degree to which epidemiological and whole-genome sequencing data agreed when describing the transmission chains of COVID-19 was required. This issue can thus be seen as an assessment of validity of the measurement of epidemiological links using epidemiological investigation.

When dealing with two categorical variables (Epicluster vs. Genomic cluster), five main indicators (Table 2) are available to estimate validity (226). While some of these are deemed more appropriate to the assessment of validity (sensitivity and specificity) and others to the evaluation of reliability (kappa or intraclass correlation), they can be used

interchangeably to evaluate validity and reliability (226). In fact, according to Szklo and Nieto, “what determines whether the evaluation is of validity or reliability is not the type of test, but what the specific goal of the evaluation is and whether a gold standard is available”. While genomic data is not a perfect measure of epidemiological links for the purposes of this work it was used as a “gold standard”. A brief review of the existing approaches is presented, providing the rationale for the selection of the analysis method.

Table 2 Summary of indices or approaches most frequently used for the assessment of validity and reliability for categorical variables. Adapted from Szklo M, Nieto JF(226).

Index or technique	**Mostly used to assess . . .**	
	Validity	Reliability
Sensitivity/specificity	++	
Youden’s J statistic	+ +	+
Percent agreement	+	+ +
Percent positive agreement	+	+ +
Kappa statistic	+	+ +

Percent agreement, percent positive agreement and kappa statistic (Table 4) are indicators to estimate the congruence between observations obtained by two different methods. These would seem the ideal here but the fact that they need a method that attributes a binary definition (that can be filled in a two-way table) makes them undoable in this case. This happens because neither epidemiological nor genomic data can provide a single classification on the cluster, this is, the cluster classification does not necessarily overlap. The same main viral genetic profile includes a high number of cases and therefore, several different epiclusters or subclusters that are not actually related among them. Table 4 shows *n* genomic and epidemiologic clusters, where only a binary variable is possible. Every single epicluster and its complement (not belonging to that epicluster) would fill cells *a*, *b*, and *c*, but *d* parcel would always be empty, since multiple different clusters could fit in that parcel. In other words, if *a* individual is belong to profile 1 and epicluster X, then *d* individuals belong to neither of those but comprehend cases that belong to profile 2,3 or 4, and cases that belong to epicluster Y or Z.

The expected tabulation between the attribution of epiclusters and main viral genetic profile is presented in Table 3. The unit of analysis is the epicluster and the variable is the number of cases in each of the main viral genetic profiles. The predominant main viral genetic profile was considered as the correct one. For example, for epicluster A, the predominant main viral genetic profile is 1, with 8 sequenced cases showing that profile. When it comes to Kappa, the non-overlapping of the two variables is a problem and supports the decision

to abandon the three previously described methods.

Table 3 Expected results using a dummy table

		Main viral genetic profile			
		1	**2**	**3**	**4**
Epicluster	**A**	8	1	0	1
	B				
	C				
	D				
	E				
	F				

Table 4 Percent agreement, percent positive agreement and Kappa statistic for paired binary test results

		Gold-standard (Main viral genetic profile)	
		1	Not 1 (2/3/4)
Method being tested (epicluster)	X	A	B
	Not X (Y/Z)	C	D

Percent agreement: $\frac{(a+d)}{(a+b+c+d)} * 100$

Positive percent agreement: $\frac{(a)}{(\frac{(a+c)+(a+b)}{2})} * 100$

Kappa statistic: $\frac{P_o-P_e}{1-P_e} = \frac{(\frac{A_o+D_o}{a+b+c+d})-(\frac{A_e+D_e}{a+b+c+d})}{1-(\frac{A_e+D_e}{a+b+c+d})}$ **(where Po is the observed agreement and Pe is the expected agreement)**

Sensitivity/specificity and Youden's J (Table 5) statistic are used to estimate the validity. Sensitivity shows the true positives and specificity true negatives. They need a clear cut-off for both categories and presuppose the existence of a gold-standard or a method that can be close to it. The J statistic is a summary index of validity that combines sensitivity and specificity. Although, theoretically, the value of J can be negative, down to –1 (in the presence of perfect disagreement, when the test always disagrees with the gold standard), a more realistic minimum for the range of J in real life is J = 0, obtained when the test performs no better than chance alone (sensitivity and specificity = 0.5) (226). This J statistic gives equal weight to sensitivity and specificity, thus assuming that both are equally important components of validity.

Table 5 Sensitivity/specificity and Youden's J for paired binary test results

		Gold-standard (Main viral genetic profile)	
		1	**Not 1 (2/3/4)**
Method being tested (epicluster)	**X**	A	B
	Not X (Y/Z)	C	D

Sensitivity: $\frac{(a)}{(a+c)}$

Specificity: $\frac{(d)}{(b+d)}$

J statistic: $\boldsymbol{Sensitivity + Specificity - 1.0}$

Considering the existing methods, a measure of sensitivity/specificity seemed more adequate. As WGS could be assumed as a gold-Standard, the question under analysis was framed as the frequency that epidemiologic data correctly identified cases within a genomic profile. Thus, sensitivity was selected.

However, this implied the calculation of several sensitivity measures (as many as the clusters obtained). A summary measure was thus useful. Metanalysis can address this need, using each epidemiologically linked cluster (epicluster) as if it were a single study.

For the metanalysis, a new database was computed, considering the epicluster as the unit of analysis, for which genotyped cases were attributed to a main genetic profile. Cases outside the predominant genetic profile were all attributed to "others" (Table 6). Epiclusters with less than 2 individuals with available genetic data were excluded from this analysis since a single individual could not be compared for congruence. Here, each epicluster was analysed as if it was a single study in a typical metanalysis; study design was not considered as a differentiating variable because the method of epidemiological investigation was the same. For this analysis, the outcome was the proportion of individuals within a single dominant genomic cluster in the total of individuals with viral WGS data, for each epicluster. A new variable, sample size, was computed considering the number of individuals with WGS data on each epicluster (3). Since participants come from a single common population and go through the same procedures performed by the same professionals under the same conditions, a fixed-effect model was considered. This assumes that observed effect sizes variation across the epiclusters is due to the random sampling error inherent to each one, namely, the within-epicluster variance (227). Nevertheless, heterogeneity in the effect sizes was assessed through τ2, I2 and the Q-statistic, to support the choice of using fixed or random effects. A forest plot was built using the logit transformation. If the proportions did not present with a normal distribution, they

were transformed into their natural logarithm for the analysis and then converted back to proportions, together with its 95% confidence intervals. To analyse the study weight, the inverse of the variance of the transformed proportion was used. Individual effect sizes were pooled and their sampling inverse variances were analysed via Linear (Fixed-Effects) Models (using the rma() function in R). The analysis used the logit transformation, since proportions higher than 80% are common and is necessary to avoid the values from exceeding 100%. Additionally, seriously misleading results have been reported using inverse of Freeman-Tukey in similar meta-analysis of single proportions, so the variance was not stabilized (228). All statistical analyses were preformed using R version 4.0.3. The level of significance was set at 5%.

Table 6 Example of the database computed to run the metanalysis

Epicluster	Predominant main genetic profile	Other	Sensitivity
A	A1	A2	A1/(A1+A2)
B	B1	B2	B1/(B1+B2)

v. Description of sporadic cases and misidentification of links

In order to resolve sporadic cases and misidentified direction of transmission, the epidemiological investigation was revised after integration of viral genomic data. This intended to inform a better understanding the hurdles and caveats associated with the epidemiological investigation during periods of high incidence rates.

A sporadic was considered, on the basis of epidemiological data, as a case that 1) did not have any data on the source of infection, either an infector or an associated cluster; or 2)for whom the source (infector or cluster) was outside the catchment area of Baixo Vouga (e.g., an infector that lived in Lisbon or a cluster in a school in Algarve). The description of these cases was based on the retrospective analysis of both data sources in order to resolve them. More detailed information was obtained on the epidemiological context, as the municipality of residence, recent travels, place of work, and the association with a setting of healthcare or nursing home institutions. Then, a genomic context was provided by a detailed inspection of phylogenetic placement of all genotyped cases into the trees reconstructed, by both 1) describing the genetic profile, not only the main viral genetic profile, and its associations with other individuals or epiclusters inside Baixo Vouga dataset, or 2) searching for similar profiles outside the Baixo Vouga dataset, by comparing it with other shared sequences in the national collection (183). Based on the previous data, a final case assessment was provided, by 1) inserting the case in a transmission chain – false

sporadic; 2) identifying the case as a singleton in both Baixo Vouga and national collection – true sporadic; or 3) declaring the case as unresolved due to the low viral genomic resolutions and high prevalence of the profile worldwide.

A misidentification refers to a link, reported by epidemiological data, that was posteriorly refuted by genomics. A misidentification could be due to the link itself between the two cases or the direction of the reported transmission. To analyse misidentifications both the ID and the hierarchical code were reported, reviewing the epicluster, the subcluster and the link context provided by epidemiological data. To assess the direction of transmission, viral genetic profiles associated with all infector-infected pairs with available genomic data were scrutinized to assess if they display a congruent ancestrality (i.e., if the epidemiologically inferred direction of transmission is supported by a hierarchically direct phylogenetic placement). Phylogenetic distance was also reported. A probable explanation was provided to the misidentification of the cases after adding pathogen genomics information, where:

1. Infector and infected could have been both infected by the same ancestral case,
2. Infector and infected could have been both infected by different ancestral cases,
3. The direction of transmission could have occurred in the other way.

4. Results

i. Context description

The first case in Baixo Vouga was on March the 8th, in Aveiro municipality (6). There was a total of 1925 confirmed COVID-19 cases from March 8th to the end of June in the region, where 39% belonged to one municipality, Ovar (6). Ovar was the first Portuguese municipality to be declared to have active community transmission (7) and, on the same day, sanitary cordon was decreed, lasting between March 17 and April 17, 2020 (8). National and local public health control measures for COVID-19 together with the epidemic curve of Baixo Vouga region are presented in Figure 8. The number of confirmed cases in Baixo Vouga peaked in the end of March and became stable and residual from the beginning of May. This timely evolution of the cases is coherent with strong measures implemented locally, nationally, and internationally.

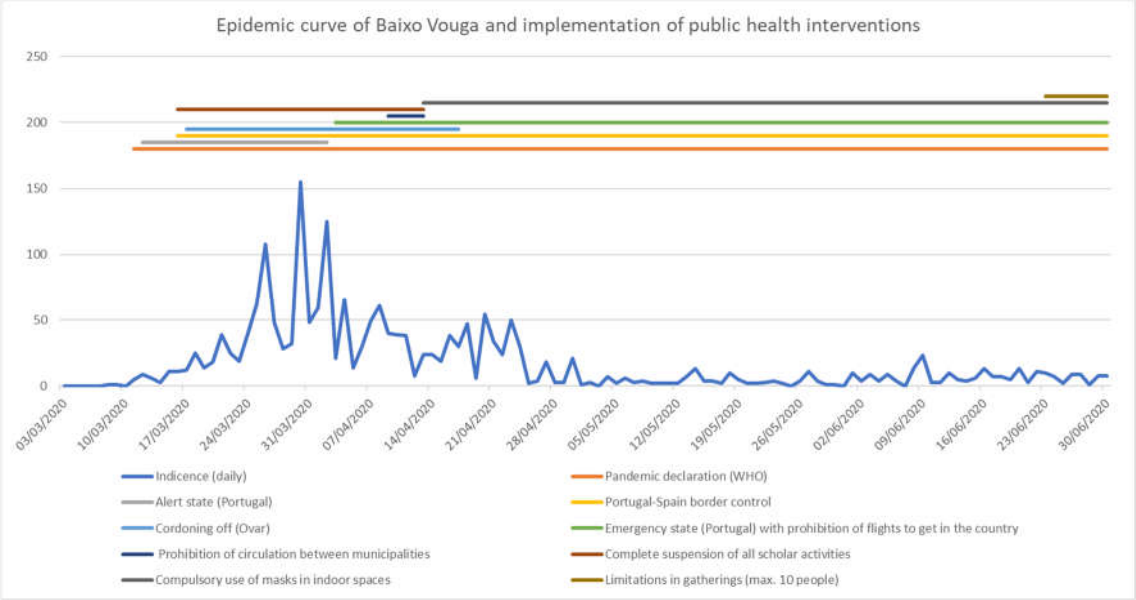

Figure 8 National and local public health control measures for COVID-19 and the epidemic curve of Baixo Vouga region. X axis – date of the diagnostic test; Y axis – Number of confirmed cases reported by Health Authorities.

ii. Study population, sample and epiclusters overview

A total of 1925 confirmed cases of COVID-19 were identified by the Baixo Vouga Public Health Unit from 8th March to 30th June 2020, based on the day of the sample collection of the diagnostic test. Of these, 62,86% were females, and the median age was 56 years (IQR 39 - 76). The proportion of individuals in the total dataset diagnosed during sanitary

cordon (according to time and geographical location) was 32,10%. Healthcare professionals accounted for 10,39% of the cases and residents in nursing homes for 15,84%. Most cases (1501, 77,97%) did not have any characterization on the context of contagion; For those whom the context was specified, the most common setting was cohabitant (51,65%), followed by work (30,66%).

From the 1143 cases assigned to epiclusters (59,38% of the total), 154 (13,47%) were considered index cases and 988 (86,44%) had an identified infector, for which 433 (43,78%) were considered direct (Table 7, Figure 9 and Appendix 2). Cases within an epicluster were slightly more frequently from a sanitary cordon area (36,66%) than sporadics ones (25,45%). Having a cohabitant as a source of contagion was verified in 15,40% within an epicluster (vs. 5,50% for sporadics). Also, certain clinical presentations were more frequent in cases within an epicluster, such as cough, myalgia, headaches, rhinorrhoea, anosmia and dysgeusia. Additionally, individuals within epiclusters were more frequently associated with closed contexts of nursing homes (21.43%) or healthcare (12.60%).

Table 7 Demographic, clinical characteristics, context of the contagion, epidemiological link type, sanitary cordon area and case settings data for Baixo Vouga COVID-19 sporadic and within-epicluster cases until 30 June 2020.

	Sporadics (N=782) n (%)*	**Cases within an epicluster (N=1143) n (%)***
Demographic		
Age (at the day of the diagnostic test) – median, IQR	55.61 [39.41;72.78]	55.53 [37.87;79.14]
Sex (Male)	315 (40.28%)	400 (35.00%)
Clinical characteristics		
Symptomatic	528 (67.52%)	736 (64.39%)
Existence of previous medical conditions	99 (37.93%)	101 (28.86%)
Context of the contagion		
Cohabitant	43 (5.50%)	176 (15.40%)
Healthcare	14 (1.79%)	19 (1.66%)
Non cohabitant acquaintances	16 (2.05%)	16 (1.40%)
Other	6 (0.77%)	4 (0.35%)
Unknown	662 (84.65%)	839 (73.40%)
Work settings	41 (5.24%)	89 (7.79%)
Epidemiological link type		
Direct	.	433 (43.78%)
Indirect	.	556 (56.22%)
Sanitary cordon (at the date of diagnosis)		
Yes	199 (25.45%)	419 (36.66%)

Case settings		
Nursing home resident or professional	347 (19.31%)	12 (9.38%)
Healthcare professional	191 (10.63%)	9 (7.03%)

*Unless indicated otherwise. IQR – Interquartile Range

When analysing the first two weeks of the epidemic, the proportion of direct links reduced over time (Figure 9), giving place to a predominance of unidentified and indirect links. This trend can be seen whenever the incidence increases. Indirect links were commonly associated with closed settings (e.g., nursing homes, healthcare facilities) were the epidemiological investigation identified the source context but not a specific infector as the epidemiological link. Such outbreaks were the origin of the two smaller peaks (end of May and beginning of June) seen in Figure 9, where a predominance of indirect link type's can be identified.

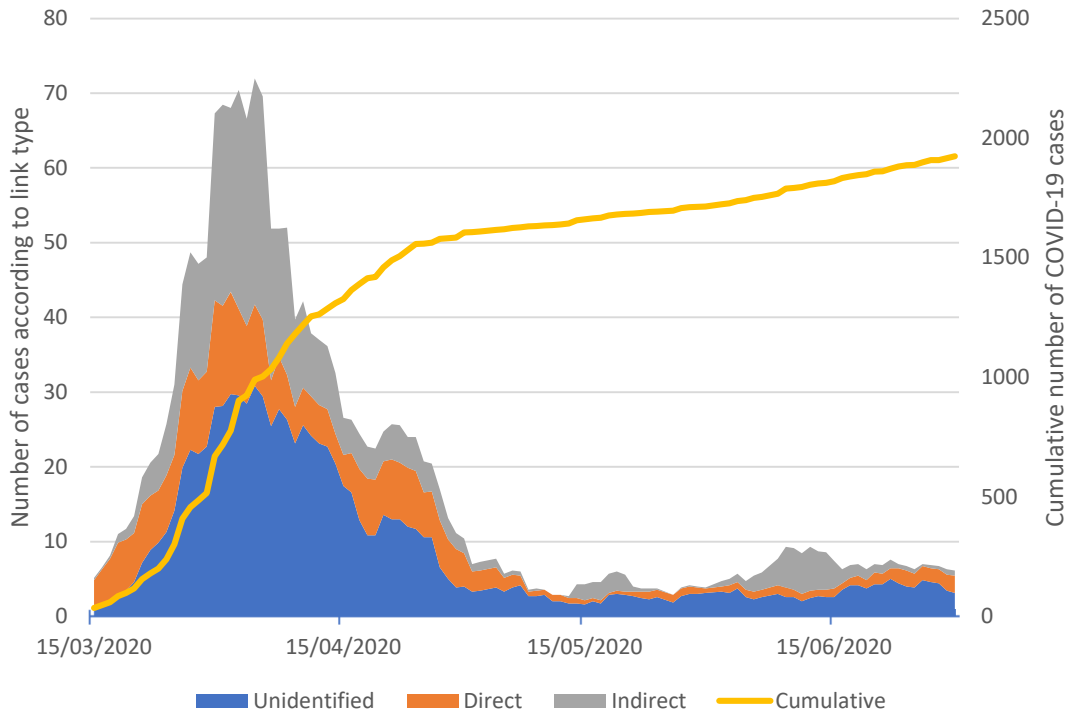

Figure 9 Weekly evolution of COVID-19 cases link type, according to the epidemiological investigation.

A total of 154 epiclusters were detected, with a median of 2 cases (IQR 2-4) (ranging from 2 to 280) (Appendix 3). A total of 10 epiclusters had more than 15 cases and the biggest included 280 (A) and 101 (B). These referred to epiclusters that spread throughout different contexts (e.g., healthcare services, nursing homes, cohabitants and other familiar contexts, work settings). Figure 10 represents the size of each epicluster, including the number of cases in each viral genetic profile, and the duration of the same epicluster (using the date

of diagnosis of the first and last case of the epicluster as limits). Most epiclusters were detected during a restricted period coincident with the sanitary cordon. After this period, epiclusters were detected less frequently and had a shorter duration Figure 10 Size, proportion of cases in each viral genetic profile, and timespan of the COVID-19 epiclusters identified in Baixo Vouga from 8th March to 30th June.

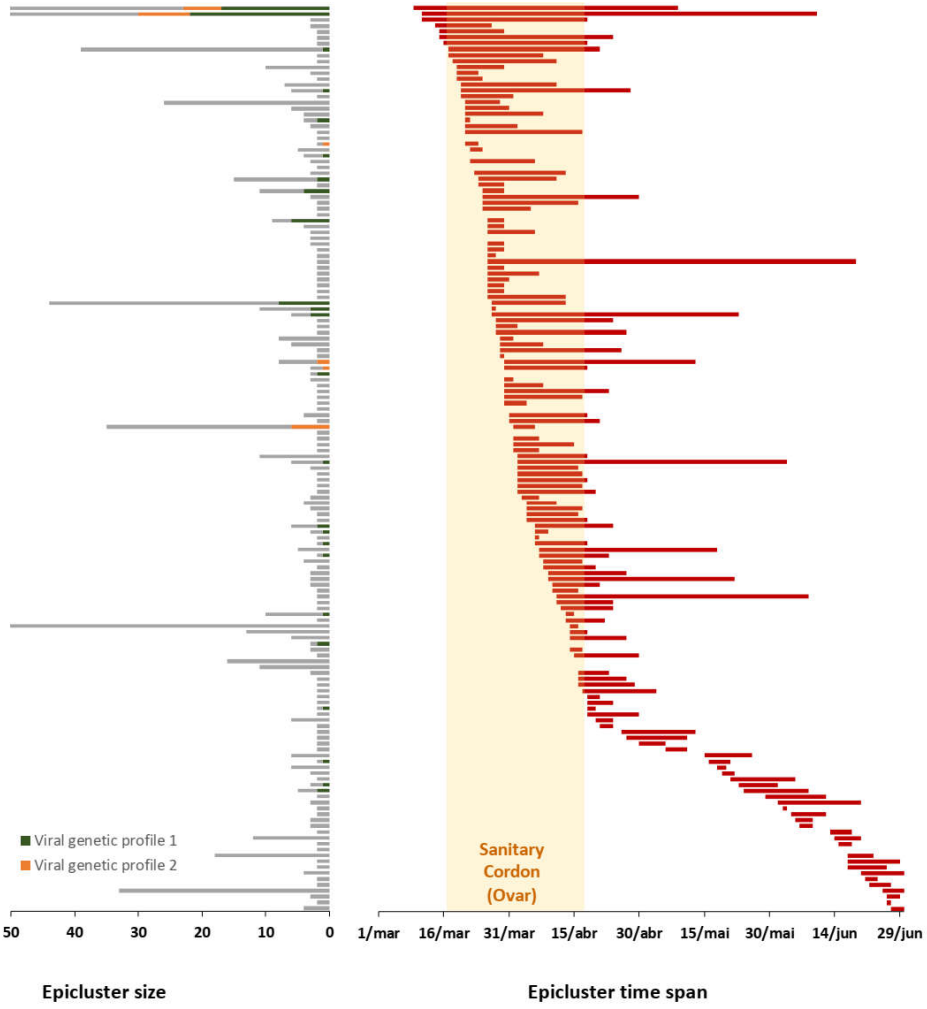

Figure 10 Size, proportion of cases in each viral genetic profile, and timespan of the COVID-19 epiclusters identified in Baixo Vouga from 8th March to 30th June.

Viral whole-genome sequencing data was available for a total of 128 cases (Appendix 4), representing 6,65% of the total dataset and 11,20% of the cases within an epicluster (Table 8). The available WGS dataset mostly covers the early epidemics, reflecting the initial focus of the national surveillance (209,229), with 127/128 samples being collected until 17 April 2020 (end of the Ovar's sanitary cordon). This subset is thus enriched by cases from the sanitary cordon area (74,22%) (Table 8) and involves a slightly higher frequency of symptomatic patients (78,12%), when compared to cases without WGS data.

Table 8 Demographic, clinical presentation, medical conditions, context of the contagion, epidemiological link type, sanitary cordon area and case settings data for Baixo Vouga COVID-19 cases with and without WGS data, until 30 June 2020.

	Cases without WGS data N=1797 n (%) *	Cases with available WGS data N=128 n (%) *	p-value
Demographic			
Age at the day of the diagnostic test – median, IQR	55.45 [38.15;75.86]	56.05 [43.44;77.21]	0.391
Sex (Male)	654 (36.39%)	61 (47.66%)	0.014
Clinical characteristics			
Symptomatic	1164 (64.77%)	100 (78.12%)	0.003
Existence of previous medical conditions	198 (32.57%)	2 (66.67%)	0.251
Epidemiological link type			
Direct	385 (43.40%)	48 (47.06%)	0.549
Indirect	502 (56.60%)	54 (52.94%)	
Sanitary cordon (at the date of diagnosis)			
Yes	523 (29.10%)	95 (74.22%)	<0.001
Case settings			
Nursing home resident or professional	347 (19.31%)	12 (9.38%)	0.008
Healthcare professional			
Healthcare professional	191 (10.63%)	9 (7.03%)	0.255
Epicluster			
Cases within an epicluster	1033 (57.48%)	110 (85.94%)	<0.001

*Unless indicated otherwise. IQR – Interquartile Range. WGS – Whole Genome Sequencing.

iii. Viral genomic diversity across epiclusters and sporadic cases

Genomic analysis of the 128 available sequences revealed four main viral genetic profiles: 95 sequences forming a sub-clade within Nextstrain clade 20A due to a shared mutation (G24077T, leading to the Spike D839Y amino acid change) (hereafter designated as genetic profile 1), 31 sequences belonging to Nextstrain clade 20B (genetic profile 2) and two genetically distant sequences belonging to Nextstrain clade 19A (genetic profiles 3 and 4) (Figure 11, Microreact interactive figure).

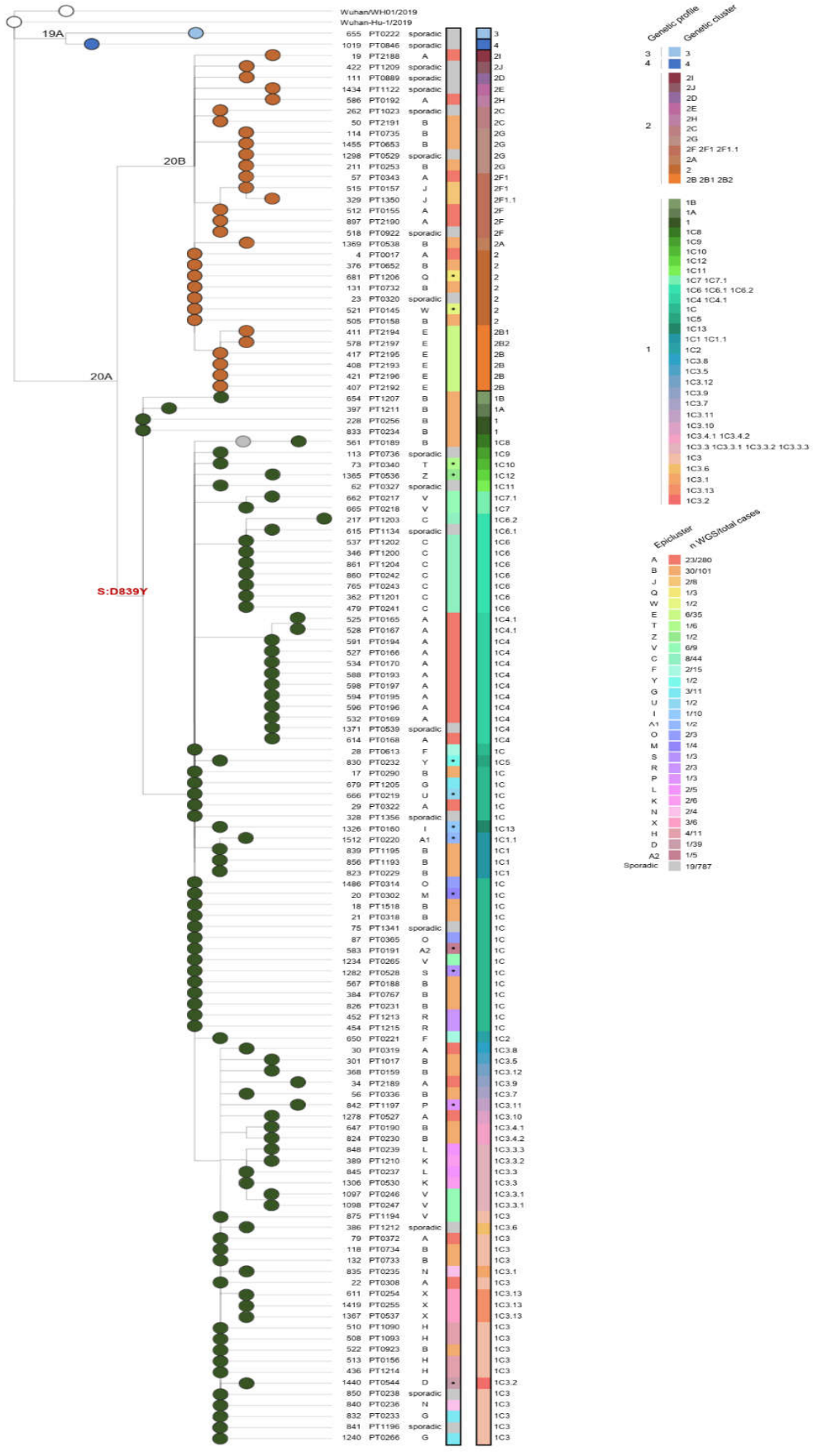

Figure 11 Phylogenetic tree of Baixo Vouga SARS-CoV-2 sequences and integration of genomic and epidemiological data. Source: INSA, 2021. Legend: 1st column: case ID; 2nd column – Viral genomic sequence code; 3rd column – epicluster; 4th column genomic cluster.

The majority of Baixo Vouga cases here identified integrate transmission chains representing ramifications of a massive dissemination of SARS-CoV-2 with genetic profile 1 (Pango lineage B.1.91) (cases represented in colour green in Figure 11). The observed genetic divergence within this transmission chains enhances the resolution power of genomics to corroborate or exclude epidemiological links. In contrast, SARS-CoV-2 with the genetic profile 2 (including Nextstrain clade 20B root sequences and 1-3 SNPs descendants) (cases represented in orange in Figure 11) was introduced multiple times, and in multiple locations, in Portugal, during the early epidemics (230). Considering the relatively low diversification of clade 20B, the power of genomics to track the source of infection and disclose direct contacts within genetic profile 2 is reduced when compared with genetic profile 1. Regarding genetic profiles 3 and 4, both represent singleton branches in the phylogenetic tree (cases represented in blue in Figure 11).

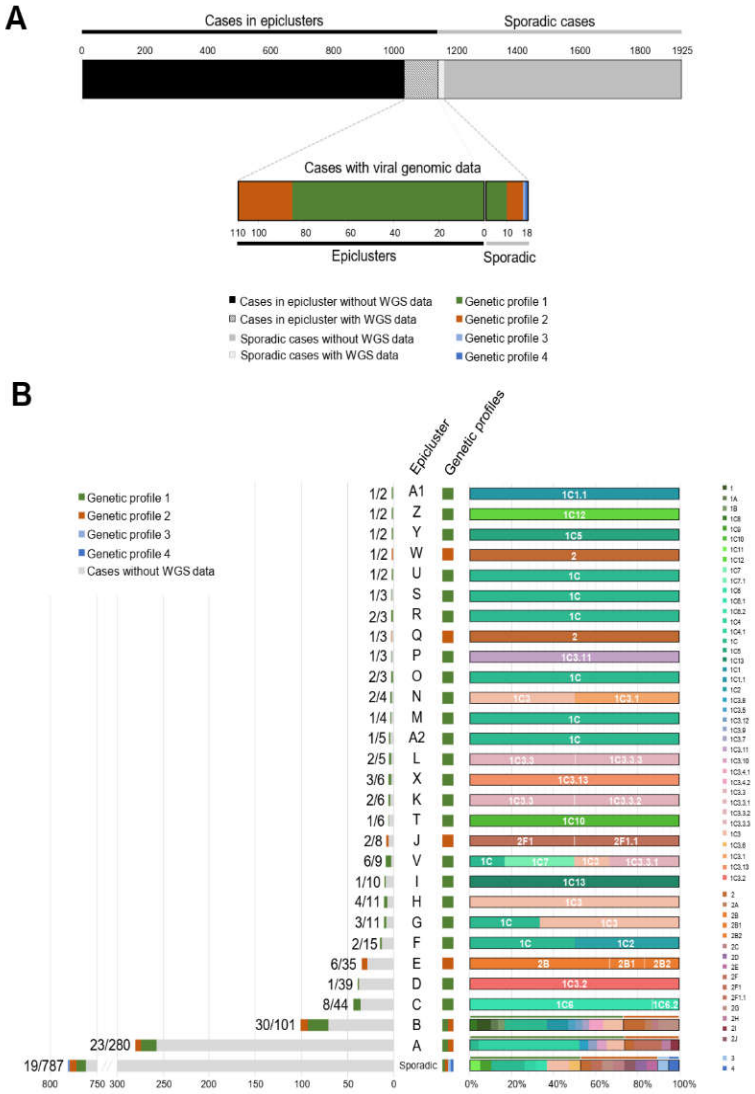

Figure 12 A – Absolute and relative frequency of cases in epiclusters and sporadics, and of cases with viral genomic data. B – Proportion of viral genetic clusters according to the epicluster. Source: INSA, 2021.

From the 128 cases with available WGS data, 18 (14,06%) were classified as sporadic and 110 (85,94%) were associated with an epicluster, according to the epidemiological investigation (Figure 12). Sporadic cases with WGS data were linked to the four main genetic profiles, including the singleton sequences of genetic profiles 3 and 4. From the 110 genotyped cases associated with epiclusters, 86 (78,18%) and 24 (21,82%) belong to genetic profiles 1 and 2, respectively. In total, there was WGS data available for cases epidemiologically linked to 28 distinct epiclusters (which together enrolled a total of 622 cases), with 15 epiclusters including more than one genotyped case (number of genotyped cases ranged from 2 to 30) (Figure 12 A).

iv. Epidemiological investigation of large epiclusters: understanding incongruences by viral genomic analysis

Genomic investigation revealed that the two largest epiclusters (A and B) mixed cases linked to two distinct SARS-CoV-2 genetic clades (here called viral genetic profiles), thus showing that each epicluster enroled more than one source/transmission network. As such, it was sought to reconstruct the transmission chains considering the integration of viral genetic data to understand the context that led to the observed incongruences.

The epidemiological investigation pointed that epicluster A, linked to healthcare and nursing home settings (Figure 13), with 280 cases (23 of which had WGS data) started with a patient that clinically presented with fever, dry cough, myalgias, headaches, and generalized weakness, that was then hospitalised. The most likely source of infection of this index case was an event that the patient attended two days prior to the symptoms' onset. Another raised possibility was a cluster in a local pharmacy, through personal relations with the workers, but this hypothesis was discarded since these cases were only diagnosed later. Indeed, the viral WGS data shows that the index case (PT0017, HC 2) had a genomic profile that was ancestral to the one found in the pharmacy cluster (PT2188, HC 2I), supporting that the index case might have been in the origin the pharmacy subcluster. The transmission chain then extended to an outpatient healthcare setting through an infected worker that attended an event with some of the pharmacy members. The viral genome data obtained from two of the patients of that healthcare service (PT0343, 2F1 and PT0192, 2H) and a close personal contact of a healthcare worker (PT2190, 2F) belonged to main genetic profile 2, thus supporting the placement in the previously described transmission chain. Contact tracing follow-up indicated that this healthcare cluster (with 36 cases) originated seven other clusters, two in nursing home settings and five in familial settings. The two familiar subclusters with available viral genomic data

belonged to the main genetic profile 1 (1C, 1C3, and 1C3.9), contradicting the epidemiological link with the healthcare subcluster. One familial cluster originated in the previously described case 2F1. The other two familial clusters with available viral genomic data originated in patients that attended the healthcare service but for whom the epidemiological link was less clear. One of these clusters was indexed to a patient (PT2189, 1C3.9), connected both to the outpatient healthcare service and a neighbouring municipality outside the Baixo Vouga region. This patient would have then infected his caregiver (PT0308 ,1C3). However, genomic analysis shows that this direction probably happened in the opposite direction, and the link to the neighbouring municipality is more consisting since other cases in that situation presented similar profiles (as seen further in this analysis). The other familial cluster was indexed to a patient (PT0322 ,1C) with two hypotheses for the epidemiological link, one in the outpatient healthcare service and the other in an inpatient service in the neighbouring municipality he had been in the previous month. Indeed, the viral WGS of the sample places the inpatient service as the most likely source. The smallest nursing home cluster presented with 85 cases, 11 of which had genome data available from the main profile 1 (1C4 (n=9) and 1C4.1 (n=2)), which supports healthcare associated transmission in this closed setting. This cluster was indexed to the domiciliary caregiver of a patient with a strong link to the outpatient healthcare service (PT0343, 2F1), who was also a worker in the nursing home. Although the nursing home worker did not have available viral WGS data, the incongruence between the genomic characteristics of the infector (main viral genetic profile 2) and the infected residents (main viral genetic profile 1, clade 20A with the additional spike mutation D839Y, dominant in the community), supports that these were two different transmission chains. The other nursing home cluster had no data on viral WGS.

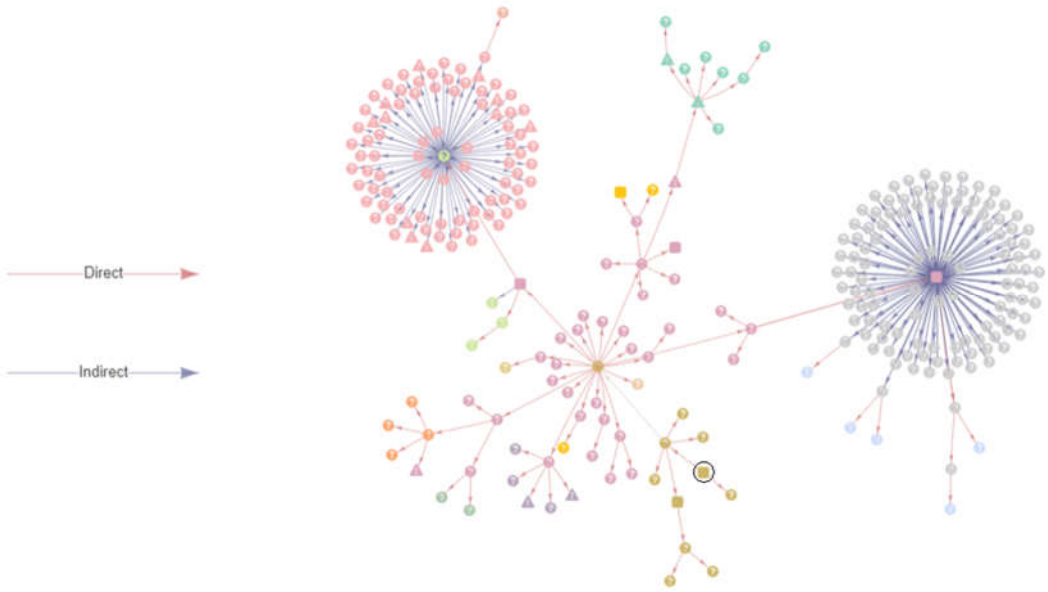

Figure 13 Epicluster A. Legend: Node Shape's: Triangle – Viral Genomic Profile 1, Square – Viral Genomic Profile 2; Node Colour's: Brown – Pharmacy cluster, Purple – Healthcare cluster, Light pink and Grey - Nursing home clusters; Circle points the index case.

Epicluster B (Figure 14), as ascertained by epidemiology, included 101 cases (30 of which had WGS data) and started with a hospitalised patient diagnosed on March 9. Although the source of infection was unknown, members of this patient's school (in a neighbouring municipality located outside the catchment are of Baixo Vouga) had returned from a trip to Italy a few days before their symptoms onset (28 February). Even though no viral genotyping data was available for this epicluster's index case, genomic data of some of its direct secondary cases belonged to main genetic profile 1 (PT0290 - 1C, through a personal contact and PT0189 - 1C8, through a contact in the school). Furthermore, the integrative analysis with the national collection of SARS-CoV-2 showed that a sequence outside the study population (PT0310, also a member of the same school) is the intermediary genome between PT0290 and PT0189, being the ancestor of PT0189. Since all these cases were associated with the main genetic profile 1 (clade 20A with the D839Y mutation in Spike), it is most likely that the index case was also infected with this SARS-CoV-2 variant. This transmission chain spread throughout cohabitants, family, friends, and healthcare workers that were in direct or indirect contact with the index case. One of these direct contacts was a healthcare worker that visited the index case, and worked in a healthcare service from the neighbouring municipality located outside the catchment area of Baixo Vouga. This healthcare service presented two clusters with ten and seven cases residing in the Baixo Vouga Region, of whom, five were sequenced (HC 1(n=2), 1C, 1C3.7, 1C3.4.1, 1C3.4.2), showing genetical coherence between them. Epidemiological

investigation was not able to place the healthcare worker as the source of this cluster. Since genomic data was not available for this case, it is not possible to disclose the direction of the transmission. Additionally, this worker had a contact with a colleague from another healthcare setting in municipality A, that also presented clusters with one sequenced case showing the main viral genetic profile 2 (HC 2C). Another healthcare cluster in municipality B, with 29 cases, presented WGS data with a mix of both main genetic profiles 1 and 2 (HC 2 (n=3), 2A, 2G (n=3), 1C3 (n=2), 1C3.5, 1C3.12). A cluster, involving six cases, in a retirement home, revealed genomic data (1C) coherent with the transmission chain of this epicluster (B). Other clusters involved a company (9 cases, linked to 1C (n=1) and 1C1 (n=3)), through a cohabitant of the index case, and a restaurant (3 cases, 1A and 1B), through a friend of the index case, with both transmissions being supported by viral genotyping.

Figure 14 Epicluster B. Legend: Node Shape's: Triangle – Viral Genomic Profile 1, Square – Viral Genomic Profile 2; Node Colour's: Green, Dark Orange, Light Orange, Pink, Grey – Hospital clusters, Brown – company; Purple - Nursing home cluster; Circle points the index case.

As observed for epicluster A, the retrospective integration of genomic data showed that the initial epidemiological reconstruction of the transmission dynamics of epicluster B was coherent with genomic data. Still, when the dimension of the epiclusters increased or extended to settings of higher population density, such as healthcare and nursing home settings (tightly connected with the community), the misidentification of epidemiological

links occurred after a certain point of the progressively more complex epidemiological investigation.

v. Sensitivity of the epidemiological investigation

Twenty six out of the 28 epiclusters (Appendix 4) were linked to a single main genetic profile (either 1 or 2), while only two (A and B) incongruently mixed cases linked to distinct main genetic profiles. This was expected since epiclusters A and B were the largest epiclusters identified, which increases the likelihood of misidentification of epidemiological links (see Epidemiological investigation of large epiclusters: understanding incongruences by viral genomic analysis). For the remaining epiclusters, the genetic profiles were highly congruent within each epicluster, i.e., each epicluster either contained 100% identical sequences or sequences with congruent ancestrality (i.e., with a hierarchically direct phylogenetic link (Figure 12 and Figure 11)).

The sensitivity of epidemiological investigation to detect links between cases that belong to the same transmission chain was high. Results of sensitivity estimated for each epicluster (with more than two cases with available viral WGS data) are provided in Table 9. Although almost all epiclusters (13/15) presented 100% of its cases within the same viral genetic profile (Table 9) there were wide confidence intervals.

Table 9 Sensitivity of the epidemiological investigation for each epicluster

Epicluster	Total number of cases per epicluster No. (genotyped)	Cases within the same main viral genetic profile (No.)	Cases outside the main viral genetic profile (No.)	Sensitivity	95%CI (lower)	95%CI (upper)
A	280 (23)	17	6	73.91%	51.59%	89.77%
B	101 (30)	22	8	73.33%	54.11%	87.72%
C	44 (8)	8	0	100.00%	63.06%	100.00%
E	35 (6)	6	0	100.00%	54.07%	100.00%
F	15 (2)	2	0	100.00%	15.81%	100.00%
G	11 (3)	3	0	100.00%	29.24%	100.00%
H	11 (4)	4	0	100.00%	39.76%	100.00%
J	8 (2)	2	0	100.00%	15.81%	100.00%
K	6 (2)	2	0	100.00%	15.81%	100.00%
L	5 (2)	2	0	100.00%	15.81%	100.00%
N	4 (2)	2	0	100.00%	15.81%	100.00%
O	3 (2)	2	0	100.00%	15.81%	100.00%
R	3 (2)	2	0	100.00%	15.81%	100.00%
V	9 (6)	6	0	100.00%	54.07%	100.00%
X	6 (3)	3	0	100.00%	29.24%	100.00%
Total	541 (97)	83	14	85.57%	93.33%	96.48%

Corresponding to epiclusters for which concordance could be assessed. i.e. more than two cases with available viral WGS data.

Pooled average sensitivity (metanalysis of proportions)

Ninety-seven (n=97, 75,78% of the total cases with viral WGS data) individuals were included in the pooled analysis of sensitivity, divided in 15 epiclusters, from which six presented a size higher than three (the median epicluster size). Low heterogeneity in the effect sizes was found ($\tau^2 = 0$; I^2 was by 0%; and the Q-statistic was 4.33 (p-value = 0.993)); for that reason, along with the fact that the methodology of epidemiological investigation was homogeneous for all epiclusters, a fixed effects model was use. The pooled sensitivity was 79.55% (95%CI: 70.42 - 86.40%). The forest plot of the obtained sensitivities for the 15 epiclusters is shown in Figure 15.

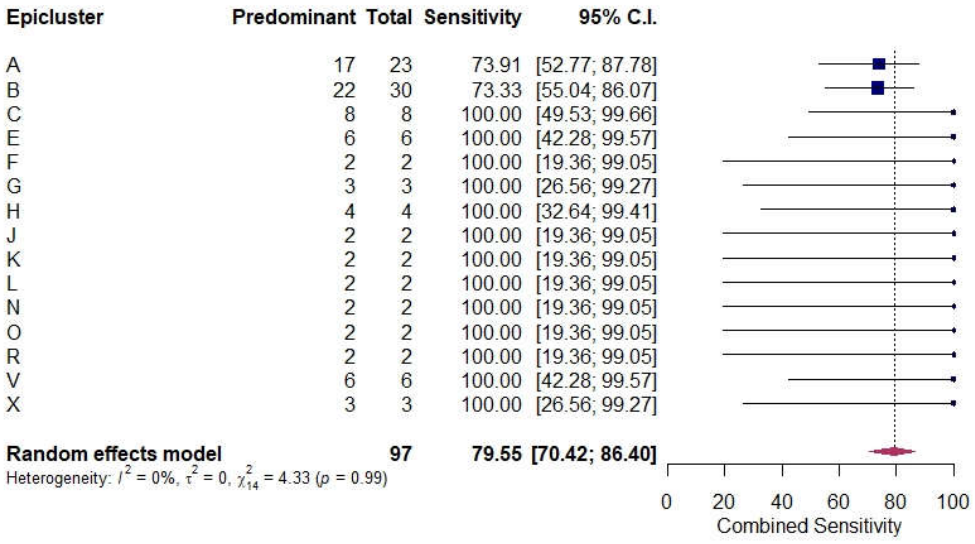

Figure 15 Forest plot with proportions' metanalysis results of epidemiological investigation sensitivity.

Legend: (Epicluster) Each epicluster with viral WGS data on more than two cases; (Predominant) The number of cases with viral WGS that belong to one predominant viral genetic profile; (Total) The total number of cases in each epicluster with viral WGS data; (Sensitivity) The proportion of the cases within the predominant viral genetic profile among the total of cases with WGS data. The CIs for individual point estimates are shown with horizontal lines. The surrounding box shows the contribution made by that individual estimate to the overall pooled estimate, weighted by the standard error of that individual series. Heterogenity (I2) = 0%. So, it used fixed effect model. Presented estimates are the inverse of the logit transformation.

vi. Review sporadic cases and misidentified direction of transmission, according to available viral whole-genome sequencing data

Table 10 summarizes the epidemiological data available for the "sporadic" cases and how genomics helped placing most of these cases within well-established transmission chains. In fact, 15 out of the 18 sporadic cases could be linked either to the largest transmission network in Baixo Vouga (driven by the SARS-CoV-2 D839Y variant, i.e., the viral genetic

profile 1), to specific epiclusters, or to COVID-19 cases identified outside the study population. The three remaining cases were either assumed as true sporadic (absence of epidemiological links or strong genomic links with any other case) or unresolved cases (lack of genomic resolution to infer a potential link to other cases).

Table 10 Cases that belonged to a viral genetic cluster but were identified as sporadic in the epidemiological investigation

Case ID (Sequence code, viral genomic profile)	Healthcare/Nursing home	Epidemiological link	Genomic context	Case assessment after genomics
1019 (PT0846, 4)	Yes	No data (case hospitalised)	The viral genetic profile is a singleton in both the Baixo Vouga dataset and the national genome collection	True sporadic case
75 (PT1341, 1C)	No	No data	Highly frequent viral genetic profile that was linked to several epicluster.	Part of the largest transmission network
328 (PT1356, 1C)	No	No data		
841 (PT1196, 1C3)	No	No data		
850 (PT0238, 1C3)	No	The epidemiological link is a cluster outside the Baixo Vouga dataset.		
65 (PT0327, 1C11)	Yes	No data (case hospitalised)	Singleton descendent of the most frequent viral genetic profile, and thus it cannot be linked to any specific epicluster.	
113 (PT0736, 1C9)	No	No data		
386 (PT1212, 1C3.6)	No	No data		
1371 (PT0539, 1C4)	No	No specific case was identified as an infector. Hypothesis: transmission occurred during the transportation of COVID-19 patients in the ambulance (part of case's work functions).	Integrates a well-defined transmission chain (1C4 and descendent profiles) within epicluster A.	
615 (PT1134, 1C6.1)	No	No data	Integrates the cluster C, which is a well-defined sub-transmission chain within epicluster A.	
23 (PT0320, 2)	No	The epidemiological link is a healthcare professional outside the Baixo Vouga dataset.	The genetic profile (Nextstrain clade 20B root) was highly prevalent worldwide, so viral genomics lack resolution to infer if this case represents a new introduction or if it integrates known transmission chains	Unresolved
262 (PT1023, 2C)	Yes	The epidemiological link is an healthcare cluster outside the Baixo Vouga dataset. Healthcare worker who was hypothetically infected by a hospitalised patient.	The genetic profile (2C) is shared by another studied case (PT2191), and both integrate a large genetic cluster including several matching sequences collected outside the dataset (mostly from neighbouring municipalities).	Linked to a transmission network beyond the geographical area
111 (PT0889, 2D)	No	The epidemiological link is a case outside the Baixo Vouga dataset, that had a history travel to Dubai.	The genetic profile (2D) is shared by another sequence collected outside the dataset (PT1470, neighbouring municipality).	
655 (PT0222, 3)	No	Travel history to Paris	The genetic profile (3) integrates a small genetic cluster including other sequences outside the dataset.	
1434 (PT1122, 2E)	No	The epidemiological link is a case outside the Baixo Vouga dataset.	The viral genetic profile (3-SNPs from clade 20B root) is a singleton in the Baixo Vouga dataset; it differs by 2-SNPs from the closest sequences (all from different regions) in the national genome collection, thus hampering the inference of a strong genetic link.	True sporadic case
518(PT0922, 2F)	No	No data	The genetic profile (2F, PT0922) is shared by other sequences of the study dataset (PT0155, PT2190); this genetic cluster, which also enrols descendant sequences (2F1, 2F1.1), represents a well-defined transmission chain, enrolling cases epidemiologically placed in epicluster A, epicluster J and another genome from the national genome database (PT0377) collected in a neighbouring municipality.	Part of a transmission network/chain within the study population
1298 (PT0529, 2G)	No	No data	The genetic profile (2G) is shared by three other studied cases from epicluster B and another genome from the national genome database (PT1352) collected in a neighbouring municipality, representing a well-defined transmission chain.	
422 (PT1209, 2J)	No	Family members outside the Baixo Vouga dataset, that had recently traveled to France.	The genetic profile (2J) is shared by another sequence collected outside the dataset (PT0250, neighbouring district).	Linked to a transmission network beyond the geographical area

Table 11 Misidentifications of the transmission between epidemiologically linked cases.

Infector - Infected pair	Epicluster	(Sub)epicluster context	Epidemiological link context	Genomic incongruence	Phylogenetic distance	Probable explanation
654 (PT1207, 1B) - 397 (PT1211, 1A)	B	Restaurant	Cohabitant	Profiles 1A and 1B are two ramifications of the same ancestral profile.	4 SNPs	The patients were infected by two distinct individuals (with ancestral or matching profiles) or by the same individual (less likely).
561 (PT0189, 1C8) - 384 (PT0767, 1C)	B	Social care institution	Cohabitant	The genomic profile of the infected (1C) is ancestral to the one from the infector (1C8).	4 SNPs	The direction of transmission occurred from 1C to 1C8, being most likely mediated by another individual(s); in fact, an intermediary genetic profile (PT0310) was found in the national genome database.
662 (PT0217, 1C7.1) - 1234 (PT0265, 1C)	V	Police station	Non-cohabitant	The genomic profile of the infected individuals is either ancestral (1C, 1C7) or represents a different ramification of the same ancestral profile (1C3, 1C3.3.1) in relation to the genetic profile collected from the infector (1C7.1).	3 SNPs	The individual with the profile 1C7.1 was most likely infected by its cohabitant (PT0218) with the profile 1C7 (congruent with symptoms onset in both patients), rather than being the infector of its cohabitants and work colleagues (as inferred by the epidemiological investigation).
662 (PT0217, 1C7.1) - 875 (PT1194, 1C3)	V	Police station	Work colleague		4 SNPs	
662 (PT0217, 1C7.1) - 1097 (PT0246, 1C3.3.1)	V	Police station	Work colleague		6 SNPs	
662 (PT0217, 1C7.1) - 1098 (PT0247, 1C3.3.1)	V	Police station	Cohabitant of case number 5		6 SNPs	
662 (PT0217, 1C7.1) - 665 (PT0218, 1C7)	V	Police station	Cohabitant		1 SNP	
34 (PT2189, 1C3.9) - 22 (PT0308, 1C3)	A	Healthcare	Healthcare/Patient/Caregiver/Family	The genomic profile of the infected (1C3) is ancestral to the one from the infector (1C3.9), differing by 3 SNPs from each other.	3 SNPs	The direction of transmission occurred from 1C3 to 1C3.9, which is congruent with the symptoms' onset of both cases and their epidemiological context in a healthcare setting (caregiver transmitted to the patient).

WGS also allowed the identification of cases for whom the inferred direction of transmission is not supported by the genomics. Reassessment of all infector-infected pairs with available genomic data enables a probable explanation for the incongruence for most cases, contextualizing them within the transmission networks taking place within the geographic area of the study (Table 11). Five incongruences found within epicluster 310, associated with a police station, were challenging to disentangle even with the integration of genomics. In fact, a single individual (PT0217), who was presumed to have infected co-habitants and work colleagues, was indeed a secondary case of one of the two co-habitants. Intriguingly, two distinct ramifications (1C3 and 1C7) of the predominant genomic profile 1, representing two distinct sub-transmission chains, were found in both settings (household and workplace). As such, the epidemiological link of PT0217 with the two settings led to the connection of all cases, jeopardizing the identification of more than one transmission chain solely by the epidemiological investigation data.

5. Discussion

Local outbreak response is highly dependent on the epidemiologic investigation to control transmission and reduce the spread of COVID-19. As such, knowing to what degree and in what contexts epidemiological investigation is failing to identify the source of infection and transmission chains is important to improve public health measures.

In this study, a retrospective integration of viral genomic data was performed in the frame of a large epidemiological investigation involving 1925 COVID-19 cases. It was carried out from March to June 2020, in Baixo Vouga Region, one of the regions with the highest incidence rates during the first epidemic wave in Portugal and the only region in the country with a municipality in sanitary cordon. Resulting from this extensive epidemiologic investigation, 154 epiclusters were detected, including 59% of the cases of the total dataset. Secondary viral genotyping data was available for 128 cases, 18 of which had been previously considered sporadic. Four main viral genetic profiles were identified, revealing 74% of samples belonging to a sub-clade within Nextstrain clade 20A due to the shared D839Y mutation in the Spike protein (profile 1), 24% sequences belonging to Nextstrain clade 20B (profile 2) and two genetically distant sequences belonging to Nextstrain clade 19A (profiles 3 and 4). These findings are coherent with the virus circulating both in Portugal and Europe in the initial phase of the epidemic (209,230,231). SARS-CoV-2 with genetic profile 1 (Pango lineage B.1.91) was most likely imported from Italy in mid-late February and spread in Portugal during the first wave. It was highly prevalent in the Northern and Central regions, being likely responsible for about 25% of all COVID-19 cases in the early epidemics in Portugal (209). As such, the Baixo Vouga cases here identified integrate transmission chains representing ramifications of this massive dissemination, which is further supported by the limited circulation of this variant in other countries. The observed genetic divergence within this transmission chains enhances the resolution power of genomics to corroborate or exclude epidemiological links. In contrast, SARS-CoV-2 with the genetic profile 2 (including Nextstrain clade 20B root sequences and 1-3 SNPs descendants) was highly prevalent worldwide (especially in Europe) during the study period (March-June 2021) (231), and was introduced multiple times, and in multiple locations, in Portugal, during the early epidemics (230). Considering the relatively low diversification of clade 20B at worldwide level, the power of genomics to track the source of infection and disclose direct contacts within genetic profile 2 is reduced when compared with genetic profile 1. Regarding genetic profiles 3 and 4, both represent singleton branches in the global national phylogenetic tree.

By analysing the epiclusters together with viral genomic data, this study allowed the identification of some hurdles and caveats associated with a large-scale investigation performed during the early spread of an emerging pandemic virus. A first remarkable observation is that two distinct SARS-CoV-2 genetic clades were mixed in the two largest epiclusters (A and B). Main viral genetic profile 1 has likely emerged in Italy and was extensively disseminated in the community in Portugal in the early epidemics (209), so it is uncertain whether the transmission network of epicluster B originated in the community or whether it reflects an additional introduction of this SARS-CoV-2 variant in Portugal (less likely). According to the epidemiological investigation, epicluster B started in the municipality of Ovar (subjected to sanitary cordon on March 17) but then extended to other territories, with only 25% of the cases belonging to the initial municipality. Epicluster A mainly contained cases mainly showing the main viral genetic profile 2, highly prevalent in Portugal and, therefore, presenting a reduced power of genomics to track the transmission chains. The retrospective integration of genomic data showed that the initial epidemiological reconstruction of the transmission dynamics of both epiclusters (A and B) was mostly accurate, with a clear dominance of the same main viral genetic profile (2 and 1, respectively).

A second remarkable observation is that likely due to the increasing dimension of the epiclusters and its extension to settings of higher population density (i.e., healthcare), the misidentification of epidemiological links inevitably occurred after a certain point. In fact, viral genomic data reveals a mix of lineages probably due to parallel transmission chains. This was most likely due to the difficulty of rapidly adapting existing resources to a high speed of viral dissemination, for which the healthcare services were not prepared, especially in the context of an epidemic of a poorly characterized infectious disease. The public health authorities logically attributed all cases inside the same closed setting to the same epicluster. However, such settings (i.e., healthcare, nursing homes) can bias the contact tracing investigation due to their high population density and tight connection with the community, thus potentiating the misidentification of epidemiological links, as previously seen in other studies of the same nature (22,28,216,217,232).

A third remarkable observation, in line with previous literature, is that this study was able to identify, inside healthcare facilities, situations of probable healthcare associated transmission (29,233–236), situations of unlikely healthcare associated transmission (237,238) and situations where two distinct transmission chains were developing in the same facility (19,28,217,232). For example, in epicluster B, a healthcare service, from a neighbouring municipality, presented two clusters with cases residing in Baixo Vouga, of

whom viral sequences (HC 1(n=2), 1C, 1C3.7, 1C3.4.1, 1C3.4.2) showed genetic coherence between them. These cases belong to the same transmission chain, although they did not directly infect each other. In particular, 1C3.4.1 and 1C3.4.2 are exclusive to this setting and hence support healthcare associated transmission. Another healthcare cluster in the same epicluster (B) presented with two different main genetic profiles, indicating the presence of parallel transmission chains. Additionally, in an outpatient healthcare setting, this study was able to confirm probable transmission, in a moment when infection prevention and control measures were not yet fully implemented in Portugal (e.g., the use of masks was only recommended for symptomatic people). Another relevant observation was the confirmation of transmission inside nursing homes, in line with previous knowledge (23,30,217,239–242). Indeed, healthcare associated transmission was supported by the clear phylogenetic segregation of 11 cases (1C4 and 1C4.1). The fact that this was a rapidly spreading cluster resulted in most samples being identical or near identical (maximum one SNP difference) (207). At the moment this cluster was detected by the local health authorities, an immediate action was developed, with the testing of all residents and staff, separation of the residents in two wings according to their viral status, and reinforcement of other measures of infection prevention and control (e.g., environmental cleaning, use of personal protective equipment's). These measures seem to have been effective since no further cases were associated to this cluster after May 2, lasting a total of 37 days. These findings are in line with previous knowledge (240) and have an important role in supporting the effectiveness of the implemented measures to prevent and control infection in closed settings, such as nursing homes. Fourthly, in company settings high genetic coherence between epidemiologically linked cases of epicluster B was identified, which may support transmission within this setting despite the implemented measures, as reinforced by the Public Health professionals subjective impressions and previous evidence (232,243–245). Fifth, this investigation illustrates a large interaction between different settings (e.g., workplace, familial, healthcare services), which is in line with previous literature (22,25,28,216,217,245).

A quantitative assessment of the epidemiological investigation validity was performed by assessing the agreement of epiclusters with the main viral genetic profiles. The overall sensitivity of epidemiological investigation to identify cases that belonged to the same transmission chain was high (70%-86%), which is in line with previous findings (22,25). While agreement regarding the main genetic profile was found for all cases within 13 out of the 15 epiclusters with at least two genotyped cases (i.e., each epicluster either contained 100% identical sequences or sequences with a direct ancestor-descendent link

at phylogenetic level), these were small clusters, thus with highly uncertain values. The lower sensitivity obtained in the pooled analysis when compared to the overall sensitivity, points the importance of combining these results to obtain a more robust estimate. To the best of my knowledge, only another study provided such quantitative estimated. It estimated a median of 100% (IQR = 93–100%) of each epidemiologically-linked cases associated with a single dominant viral genetic cluster (22). While the results are similar, the differences in sample size, context, methodology and low viral genetic diversity observed during the first months of SARS-CoV-2 expansion in human population, challenges a direct comparison.

The introduction of cases into closed, highly populated, settings and the high COVID-19 incidence might explain an imperfect sensitivity to identified links between cases using epidemiological data. An increased proportion of indirect or unknown links was identified when the incidence increases or when in presence of outbreaks in closed settings (i.e., healthcare and nursing homes). Indirect links are a gross approximation of the source of infection using the cluster. They are attributed when the EI was not able to find and individual case as the link. Attributing an indirect link increases the chance of having a case falsely associated to an epicluster. In a context of community transmission and rapid spread, these findings suggest a decrease in the reliability of the epidemiological links over time, which is also in line with previous literature (216). In fact, progressively more complex epidemiological investigation provided by such conditions could explain the lower sensitivity of our findings, compared to other studies that reported sensitivity in the general population (22). This is in line with a previous quantitative analysis taken exclusively in these settings, that showed a 45% sensitivity of epidemiological investigation, using viral genomic data (216).

Overall, these results suggest the need to use research methods that can adjust for the reduction on the sensitivity in situations such as closed settings and high incidence. For example, some solutions could be the imputation of the epidemiological links or use of multivariate analysis where closed settings and high incidence periods work as confounders. For the practical utility in the field, when in situations of outbreaks in closed settings and high incidence periods, the increase of resources or the use of tools that can respond in an immediate and reliable way could present as solutions to maintain the levels of sensitivity.

In this retrospective investigation, genomics allowed to understand some sporadic cases and misidentified infector-infected direction of transmission. Although only a few cases

could be screened, they were very informative at several levels by unravelling distinct contexts where epidemiological investigation was limited either because not enough information could be collected or due to the lack of resolution to the links and/or the direction of transmission.

For instance, genomics allowed 13 “sporadic” cases to be linked to Baixo Vouga transmission chains, while others (n=5) were associated with cases and/or genetic clusters expanding beyond the study region (as assessed by the additional integration of the viral data into the nationwide genome collection). In fact, this study shows the inclusion of “sporadic” cases into well-defined genetic clusters / epiclusters (e.g., PT1134 fell within 1C6 genetic cluster, which perfectly overlapped with epicluster 95; PT0539 revealed the genetic profile 1C4, which is shared by other nine sequences, all of them linked to epicluster 3).

In what comes to misidentified directions of infection, this study found a peculiar situation between patient and caregiver, in which the WGS data allowed the identification of the transmission as being from the second to the first. Also, a cluster including members from a police station (1C7.1), where a worker was assumed as the infector of both colleagues and its cohabitants, has shown to have been misidentified. In fact, a genotyped cohabitant presented a more ancestral profile (1C7), indicating a different direction of transmission. Intriguingly, two distinct ramifications (1C3 and 1C7) of the main viral genomic profile 1, representing two distinct sub-transmission chains were found, which could be consistent with two different introductions in this setting. Although some intermediary cases are not represented in our dataset, two colleagues (1C3 and 1C3.3.1.) most certainly belong to the same transmission chain, along with two more (1C) that are less clear. Apart from the difficulties that are inherent to the epidemiological investigation (i.e., collected data is highly dependent on the information provided by the patient), these observations indicate the need of having robust surveillance systems integrated with pathogen genomic sequencing.

Additional gains were achieved with the integrative approach of this study. The identification of matching sequences across several epiclusters supports that they are most likely part of the same transmission network (e.g. genetic profile 1C was detected in patients linked to 10 distinct epiclusters and in 3 sporadic cases). Also, the identification of epiclusters or individual cases, potentially linked to epiclusters that extended beyond the studied population, was achieved by the integration of the studied sequences in the frame of the national SARS-CoV-2 genetic diversity (e.g., PT0232, which was linked to the 2-case epicluster 371, integrated a large genetic cluster in the global phylogenetic tree) (246).

In sum, the addition of pathogen genomic sequencing allowed the understanding of sporadics and misidentifications, the connection between several epiclusters (probably reflecting a wider transmission chain), and the insertion of Baixo Vouga epiclusters and individual cases within the national transmission chains, suggesting transmission between geographical areas. International health authorities are strongly promoting the urge integration of genome sequencing data as an essential component of surveillance systems (247). In Portugal, such framework is partially established, but still needs to be strengthened to automate the systematic integration of genomic and epidemiological data, and to promote timely communication of results back to local public health entities (230). These efforts will not only require the re-design of currently implemented systems towards a more comprehensive and interactive data flow between all stakeholders, but also require that all public health teams and decision-makers are aligned with the need to embed genomics into the routine activities of infectious disease surveillance and outbreak resolution. The COVID-19 pandemic highlighted this field by establishing genomic surveillance as a key tool for Public Health decision-making towards outbreak control. Still, despite the great advances in this field for SARS-CoV-2, there are huge discrepancies observed between countries in the establishment of frameworks for integration of epidemiological and genomic data (210–212,212–214,248). In fact, genomics-informed public health outbreak responses at country levels theoretically demand multiple resources and logistics, involving wide participation of laboratories, (de)centralized data analysis (depending on the geographical organization of the health system: national, regional, or local), and the rapid sharing of integrated genomic and epidemiological data. This study was able to show the practical utility of such tools and a framework for its implementation. The information obtained offer an additional value to support local/regional epidemiological investigation, outbreak resolution and evaluation of non-pharmacological interventions, such as contact-tracing or cordon-sanitaire, this study and others (22,23,25,210,216,217,248,249) were able to show. The replication of this framework is considered possible in the entire Portuguese territory since the national reference lab, the surveillance tools and the core capacities of Public Health professionals are resemblant. A factor that could limit its implementation could be associated with the local receptiveness to a new approach.

This study presentes some limitations. First, the subset with viral genomic data included disproportionally more residents from Ovar (sanitary cordon area) suggesting a potential selection bias. The public health “focus” on the sanitary cordon area likely justifies the overrepresentation of this kind of samples in the national genome collection (from where

sequence data was retrospectively retrieved). This bias could have influenced the sensitivity in two directions. On the one hand, leading to a reduced sensitivity, due to higher incidence and more frequency of outbreaks in closed settings. Also, since community transmission was soon declared in Ovar and many cases were attributed as an epidemiological link residing in that area, and no further investigation was taken to find the specific infector or cluster. On the other hand, the excess of cases from the sanitary cordon area could have increased the sensitivity, due to a higher focus of health authorities in that region, therefore allocating more resources (such as Public Health professionals to preform epidemiological investigation, test capacity and viral WGS). Second, the subset with viral genomic data presented more frequently symptomatic cases, which could suggest a selection bias, due to the method of sample selection to perform WGS. In fact, higher sequencing success is obtained for samples with higher viral load, which correlates with symptomatology (250). This could affect the sensitivity estimates by increasing them, since asymptomatic cases are usually more difficult to track and, therefore, more commonly associated with an unknown or misidentified source of infection. Also, asymptomatic cases could mislead their own contact tracing for the fact that the infectious period in these individuals is arbitrary and less precise.

The estimates of sensitivity of this study might be influenced by the quality of data collected during EI leading to information bias (226). When the information is obtained during EI these estimates can vary according to the characteristics of the individuals in question (e.g., the importance that each case or PHU staff gave to a specific aspect, or how well a case remembered its contacts in the past two weeks), therefore, reducing the reliability of the collected epidemiological data. Also, the data from the EI was not collected for research purposes but for local outbreak control and communication with other levels of services. This leads to records that are not tightly standardized and often lack relevant information. These aspects could lead to a misclassification of the epidemiological link, therefore reducing the sensitivity of the EI.

Finally, phylogenetic trees provide a reconstruction of similarity among viral WGS according to genome differences (SNPs) related to a common ancestor that depend on the dataset and on pathogen intrinsic characteristics (e.g., the amount of time it has been introduced in the human population, the mutation rate). In this study, the small amount of time since the SARS-CoV-2- virus introduction in human population led to a relatively low genetic diversity and therefore highly similar genomes. As such, the option to use main genetic profiles, instead of a cut-off with higher resolution, was taken to guarantee that cases associated to distinct genetic profiles could not belong to the same epicluster. In

alternative, if another cut-off that allowed a higher power of resolution would have been chosen it would likely decrease the sensitivity. The dependence of sensitivity estimates on the cut-off level (i.e., attribution to four different main viral genetic profiles) used shows that this is a variable to be considered and deeply examined when conducting such type of concordance studies. However, in this study this bias is very unlikely since we used main viral genetic profiles to ascertain a transmission chain. Future investigation referent to later stages of the COVID-19 epidemics can shed further light in this matter.

The small sample size results in statistical imprecision, by the increased probability of random errors. In fact, this study presents a relatively low proportion of sequenced viral samples (6.5%) from the total study population when compared to similar research (22,23,25,28). This is a direct consequence of the study involving a retrospective integration of the available SARS-CoV-2 national sequence database, rather than involving a "real-time" sequencing.

In the quantitative analysis, the 15 epiclusters had a small number of sequenced cases, thus associated with highly uncertain values. However, by performing a metanalysis the sample size was weighted in the final estimates, thus overcoming this issue. The metanalysis performed used fixed effect model. While this tends to be inferior to a random-effects model (251), here epidemiological investigation (EI) was performed in the same region, by the same professionals, and following the same guidelines. So, the effect across epiclusters is not likely to be heterogeneous, as is usual in medical research studies, which are typically conducted using different protocols and in different contexts.

The Portuguese epidemiological surveillance system and associated procedures are homogenously implemented at the national level. Therefore, we expect that the resources, context, and challenges associated with the epidemiological investigation in other regions, as well as the main hurdles, are possibly similar to the ones presented here for Baixo Vouga. Even though, previous capacity building and some specific actions taken by the Public Health Unit of Baixo Vouga might have facilitated the ability to conduct such massive investigation involving hundreds of epidemiological inquiries. Harmonized data collection and management, teams fully dedicated to epidemiological investigation and contact tracing, training and automatization of processes, and the ability of rapidly mobilising trained professionals according to the necessities are examples of such actions. For instance, a previously created internal tool, allowing the integration of rich epidemiological data and a more comprehensive linkage of cases, turned out to be critical to better follow transmission chains. Furthermore, professionals and information technology tools (with the

support of the management bodies) were rapidly mobilized to support the investigation, also enhancing the capacity to perform timely contact tracing. Such improvements could be limiting the generalisability of our finding to other contexts. Also, time-specific circumstances, as the sanitary cordon, were key for the mobilisation of resources and could limit the comparison of our estimates in other periods considering the same geographical area.

Key sociodemographic variables play a key role on the estimates of sensitivity and present as a problem related to their extrapolation to populations with different characteristics. For example, we may hypothesize that the EI is less sensitive for young people when comparing to elderly, since they are normally less compliant and present with more diverse and numerous social contacts (252,253). The sample used in this study was older than the Portuguese population (median age 55, versus 46)(7) which would probably overestimate sensitivity and also limit generalisability to other settings. A stratified analysis would be of great utility, but the small sample makes it a challenge.

6. Conclusions

This research was able to describe in detail COVID-19 transmission chains with great dimension, in an early phase of the epidemic, including a setting of sanitary cordon, in Portugal. Epidemiological reconstruction of the transmission chains presented high sensitivity. However, high incidence and closed settings led to misidentification of links. Difficulties associated with the epidemiological investigation were identified, in the assessment of hundreds of community cases in the context of a massive epidemics and a sanitary cordon. This research was not only able to assess the reliability of epidemiological methods in identifying the transmission chains of COVID-19, but also allowed the uncovering of previously unknown links. Additionally, this research provides a framework for future systems, integrating Pathogen Genomics in field epidemiology. It constitutes an example of how the integration of viral WGS in epidemiological approaches can strongly support and complement the tracking of pathogen spread and, therefore, give insights on the effectiveness of implemented outbreak control measures, guiding preparedness actions to face future epidemics.

Recommendations

Future studies in this field are needed to clarify some details not fully explained, such as the apparent decrease of sensitivity in healthcare or nursing home settings. Additionally, bigger sample sizes could be helpful in obtaining more robust information. Finally, instead of a retrospective integration of viral WGS data with epidemiological data, a prospective analysis of cases, with random selection of genotypes samples, could be more informative and could overcome some of the representativeness issues and selection bias.

In terms of working framework, a systematic and real-time integration of viral WGS and epidemiological data could also help local health authorities understand the transmission dynamics and the effectiveness of public health measures. The installation of surveillance and information systems, capable of containing all demographic, clinical and epidemiological data (including the storage of the infector-infected duplet that derives from the epidemiological investigation in an intelligible way that can allow the construction of epiclusters) and of merging pathogen WGS data in real time, is recommended. These tools should be bidirectionally informative (i.e., bottom-top and top-bottom) and reach all levels of health authorities and laboratorial partners. Such features of an integrative surveillance system, by contributing to timely and evidence-informed political decision-making in public health, have the potential to decrease the burden of infectious diseases.

7. Bibliographic references

1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med [Internet]. 24 de Janeiro de 2020 [citado 2 de Junho de 2020]; Disponível em: https://www.nejm.org/doi/10.1056/NEJMoa2001017

3. World Health Organization. IHR Emergency Committee on Novel Coronavirus (2019-nCoV) [Internet]. 2020 [citado 3 de Junho de 2020]. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov)

4. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 [Internet]. 2020 [citado 3 de Junho de 2020]. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

5. Direção Geral da Saúde. Novo Coronavírus COVID-19: Relatório da Situação Epidemiológica em Portugal. Lisboa; 2020 Jul.

6. Unidade de Saúde Pública do ACES Baixo Vouga. Relatório de situação da COVID-19 na região do Baixo Vouga. Aveiro; 2020 Jun.

7. National Institute of Statistics. Resident population (No.) by Place of residence (Ovar, NUTS III - 2013), Annual, 2018 [Internet]. Statistics Portugal. 2020 [citado 3 de Junho de 2020]. Disponível em: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0008273&selTab=tab0

8. Resolution of the Council of Ministers 10-D / 2020, 2020-03-19 [Internet]. Electronic Journal of the Republic. [citado 3 de Junho de 2020]. Disponível em: https://dre.pt/web/guest/home/-/dre/130413790/details/maximized?serie=I&day=2020-03-19&date=2020-03-01

9. Global Initiative on Sharing All Influenza Data (GISAID) [Internet]. [citado 20 de Maio de 2021]. Disponível em: https://www.gisaid.org/

10. Nextstrain [Internet]. [citado 23 de Janeiro de 2021]. Disponível em: https://nextstrain.org/

11. Stefanelli P, Faggioni G, Presti AL, Fiore S, Marchi A, Benedetti E, et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. Eurosurveillance. 2 de Abril de 2020;25(13):2000305.

12. Marquez S, Prado-Vivar B, Guadalupe JJ, Gutierrez Granja B, Jibaja M, Tobar M, et al. Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19 in Ecuador. medRxiv. 14 de Junho de 2020;

13. Harilal D, Ramaswamy S, Loney T, Suwaidi HA, Khansaheb H, Alkhaja A, et al. SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission. Clin Chem. 1 de Novembro de 2020;66(11):1450–8.

14. Xavier J, Giovanetti M, Adelino T, Fonseca V, Barbosa da Costa AV, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. Emerg Microbes Infect. Dezembro de 2020;9(1):1824–34.

15. Sekizuka T, Itokawa K, Hashino M, Kawano-Sugaya T, Tanaka R, Yatsu K, et al. A Genome Epidemiological Study of SARS-CoV-2 Introduction into Japan. mSphere. 11 de Novembro de 2020;5(6).

16. Maurano MT, Ramaswami S, Zappile P, Dimartino D, Boytard L, Ribeiro-Dos-Santos AM, et al. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. Genome Res. Dezembro de 2020;30(12):1781–8.

17. Walker A, Houwaart T, Wienemann T, Vasconcelos MK, Strelow D, Senff T, et al. Genetic structure of SARS-CoV-2 reflects clonal superspreading and multiple independent introduction events, North-Rhine Westphalia, Germany, February and March 2020. Euro Surveill. Junho de 2020;25(22).

18. Handrick S, Bestehorn-Willmann M, Eckstein S, Walter MC, Antwerpen MH, Naija H,

et al. Whole genome sequencing and phylogenetic classification of Tunisian SARS-CoV-2 strains from patients of the Military Hospital in Tunis. Virus Genes. Dezembro de 2020;56(6):767–71.

19. Sikkema RS, Pas SD, Nieuwenhuijse DF, O'Toole Á, Verweij J, van der Linden A, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. Lancet Infect Dis. Novembro de 2020;20(11):1273–80.

20. Thielen PM, Wohl S, Mehoke T, Ramakrishnan S, Kirsche M, Falade-Nwulia O, et al. Genomic Diversity of SARS-CoV-2 During Early Introduction into the United States National Capital Region. medRxiv. 15 de Agosto de 2020;

21. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. PNAS. 28 de Abril de 2020;117(17):9241–3.

22. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nature Communications. 1 de Setembro de 2020;11(1):4376.

23. da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, et al. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. Nat Microbiol. Janeiro de 2021;6(1):112–22.

24. Puenpa J, Suwannakarn K, Chansaenroj J, Nilyanimit P, Yorsaeng R, Auphimai C, et al. Molecular epidemiology of the first wave of severe acute respiratory syndrome coronavirus 2 infection in Thailand in 2020. Sci Rep [Internet]. 6 de Outubro de 2020 [citado 26 de Abril de 2021];10. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7538975/

25. Al-Mahruqi S, Al-Wahaibi A, Khan AL, Al-Jardani A, Asaf S, Alkindi H, et al. Molecular epidemiology of COVID-19 in Oman: A molecular and surveillance study for the early transmission of COVID-19 in the country. Int J Infect Dis. Março de 2021;104:139–49.

26. Di Giallonardo F, Duchene S, Puglia I, Curini V, Profeta F, Cammà C, et al. Genomic Epidemiology of the First Wave of SARS-CoV-2 in Italy. Viruses. 14 de Dezembro de 2020;12(12).

27. Böhmer MM, Buchholz U, Corman VM, Hoch M, Katz K, Marosevic DV, et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-

associated primary case: a case series. The Lancet Infectious Diseases. Agosto de 2020;20(8):920–8.

28. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. The Lancet Infectious Diseases. 1 de Novembro de 2020;20(11):1263–72.

29. Paltansing S, Sikkema RS, de Man SJ, Koopmans MPG, Oude Munnink BB, de Man P. Transmission of SARS-CoV-2 among healthcare workers and patients in a teaching hospital in the Netherlands confirmed by whole-genome sequencing. J Hosp Infect. Abril de 2021;110:178–83.

30. Lucey M, Macori G, Mullane N, Sutton-Fitzpatrick U, Gonzalez G, Coughlan S, et al. Whole-genome Sequencing to Track Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Transmission in Nosocomial Outbreaks. Clin Infect Dis. 1 de Junho de 2021;72(11):e727–35.

31. John M. Last. A dictionary of epidemiology. 4th edition. 2001.

32. Johan Giesecke. Modern Infectious Disease Epidemiology. Third edition. 2017.

33. Disease factsheet about tetanus [Internet]. European Centre for Disease Prevention and Control. [citado 16 de Novembro de 2021]. Disponível em: https://www.ecdc.europa.eu/en/tetanus/facts

34. Leon Gordis. Epidemiology. 2014. (5th edition).

35. Guest C, Ricciardi W, Kawachi I, Lang I. Oxford Handbook of Public Health Practice - Oxford Medicine. 3 edn. Oxford University Press; 2013.

36. CDC. Principles of Epidemiology in Public Health Practice. An Introduction to Applied Epidemiology and Biostatistics. Third Edition. 2011.

37. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Lippincott Williams & Wilkins; 2008. 776 p.

38. GBD Compare | IHME Viz Hub [Internet]. [citado 5 de Abril de 2021]. Disponível em: http://vizhub.healthdata.org/gbd-compare

39. WHO | What are the health risks related to overcrowding? [Internet]. WHO. World Health Organization; [citado 19 de Abril de 2021]. Disponível em: http://www.who.int/water_sanitation_health/emergencies/qa/emergencies_qa9/en/

40. Roger Detels, Martin Gulliford, Quarraisha Abdool Karim, Chorh Chuan Tan. Global Public Health. 6th edition. Oxford University Press; 2015.

41. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 17 de Outubro de 2020;396(10258):1204–22.

42. Omran AR. The Epidemiologic Transition: A Theory of the Epidemiology of Population Change. Milbank Q. Dezembro de 2005;83(4):731–57.

43. Barbosa H, Godinho A. Crises de mortalidade em Portugal desde meados do século XVI até ao início do século XX [Internet]. Universidade do Minho. Núcleo de Estudos de População e Sociedade (NEPS); 2001 [citado 19 de Abril de 2021]. Disponível em: http://repositorium.sdum.uminho.pt/

44. Condran GA, Crimmins-Gardner E. Public health measures and mortality in U.S. cities in the late nineteenth century. Hum Ecol. Março de 1978;6(1):27–54.

45. Vaccine-preventable diseases [Internet]. European Centre for Disease Prevention and Control. [citado 5 de Abril de 2021]. Disponível em: https://www.ecdc.europa.eu/en/immunisation-vaccines/facts/vaccine-preventable-diseases

46. J. Larry Jameson, Anthony S. Fauci, Dennis L. Kasper, Stephen L. Hauser, Dan L. Longo, Joseph Loscalzo. Harrison's Principles of Internal Medicine, 20e [Internet]. 2018 [citado 29 de Abril de 2021]. Disponível em: https://accessmedicine.mhmedical.com/book.aspx?bookID=2129

47. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature. Novembro de 2020;587(7834):448–54.

48. Antibiotic resistance [Internet]. [citado 19 de Abril de 2021]. Disponível em: https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance

49. Ten health issues WHO will tackle this year [Internet]. [citado 19 de Abril de 2021]. Disponível em: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

50. Nextstrain / ncov / gisaid / global [Internet]. [citado 16 de Novembro de 2021]. Disponível em: https://nextstrain.org/ncov/gisaid/global?l=clock

51. Neutrality and Molecular Clocks | Learn Science at Scitable [Internet]. [citado 16 de Novembro de 2021]. Disponível em: https://www.nature.com/scitable/knowledge/library/neutrality-and-molecular-clocks-100492542/

52. Spiteri G, Fielding J, Diercke M, Campese C, Enouf V, Gaymard A, et al. First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. Euro Surveill. 5 de Março de 2020;25(9):2000178.

53. World Health Organisation. International health Regulations. 2005.

54. Weekly epidemiological update - 29 December 2020 [Internet]. [citado 17 de Janeiro de 2021]. Disponível em: https://www.who.int/publications/m/item/weekly-epidemiological-update---29-december-2020

55. Directorate-General of Health - DGS. COVID-19 Situation Report - 300 [Internet]. 2020 Dez. Report No.: 300. Disponível em: https://covid19.min-saude.pt/wp-content/uploads/2020/12/300_DGS_boletim_20201227.pdf

56. Diário da República. Despacho n.º 3372-C/2020 | DRE [Internet]. Diário da República Eletrónico. [citado 16 de Novembro de 2021]. Disponível em: https://dre.pt/dre/detalhe/despacho/3372-c-2020-130366471

57. Diário da República. Resolução do Conselho de Ministros n.º 18-B/2020 [Internet]. Diário da República Eletrónico. [citado 16 de Novembro de 2021]. Disponível em: https://dre.pt/

58. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 17 de Março de 2020;323(11):1061–9.

59. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical Characteristics of Coronavirus

Disease 2019 in China. N Engl J Med [Internet]. 28 de Fevereiro de 2020 [citado 18 de Junho de 2020]; Disponível em: https://www.nejm.org/doi/10.1056/NEJMoa2002032

60. Baj J, Karakuła-Juchnowicz H, Teresiński G, Buszewicz G, Ciesielka M, Sitarz E, et al. COVID-19: Specific and Non-Specific Clinical Manifestations and Symptoms: The Current State of Knowledge. Journal of Clinical Medicine. Junho de 2020;9(6):1753.

61. Garg RK. Spectrum of Neurological Manifestations in Covid-19: A Review. Neurol India. Junho de 2020;68(3):560–72.

62. Sachdeva M, Gianotti R, Shah M, Lucia B, Tosi D, Veraldi S, et al. Cutaneous manifestations of COVID-19: Report of three cases and a review of literature. J Dermatol Sci. 29 de Abril de 2020;

63. The Prevalence of Olfactory and Gustatory Dysfunction in COVID-19 Patients: A Systematic Review and Meta-analysis - Jane Y. Tong, Amanda Wong, Daniel Zhu, Judd H. Fastenberg, Tristan Tham, 2020 [Internet]. [citado 10 de Janeiro de 2021]. Disponível em: https://journals.sagepub.com/doi/full/10.1177/0194599820926473

64. Lee Y, Min P, Lee S, Kim S-W. Prevalence and Duration of Acute Loss of Smell or Taste in COVID-19 Patients. Journal of Korean Medical Science [Internet]. 3 de Março de 2020 [citado 10 de Janeiro de 2021];35(18). Disponível em: https://doi.org/10.3346/jkms.2020.35.e174

65. Utility of hyposmia and hypogeusia for the diagnosis of COVID-19 - The Lancet Infectious Diseases [Internet]. [citado 10 de Janeiro de 2021]. Disponível em: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30297-8/fulltext

66. Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodriguez A, et al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. Eur Arch Otorhinolaryngol. 6 de Abril de 2020;

67. Hornuss D, Lange B, Schröter N, Rieg S, Kern WV, Wagner D. Anosmia in COVID-19 patients. Clin Microbiol Infect. Outubro de 2020;26(10):1426–7.

68. Patel A, Charani E, Ariyanayagam D, Abdulaal A, Denny SJ, Mughal N, et al. New-onset anosmia and ageusia in adult patients diagnosed with SARS-CoV-2 infection.

Clinical Microbiology and Infection. 1 de Setembro de 2020;26(9):1236–41.

69. A review of smell and taste dysfunction in COVID-19 patients - PubMed [Internet]. [citado 10 de Janeiro de 2021]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32918429/

70. Alimohamadi Y, Tola HH, Abbasi-Ghahramanloo A, Janani M, Sepandi M. Case fatality rate of COVID-19: a systematic review and meta-analysis. J Prev Med Hyg. Junho de 2021;62(2):E311–20.

71. Treskova-Schwarzbach M, Haas L, Reda S, Pilic A, Borodova A, Karimi K, et al. Pre-existing health conditions and severe COVID-19 outcomes: an umbrella review approach and meta-analysis of global evidence. BMC Med. 27 de Agosto de 2021;19(1):212.

72. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 7 de Abril de 2020;323(13):1239.

73. Criteria for releasing COVID-19 patients from isolation [Internet]. [citado 10 de Maio de 2021]. Disponível em: https://www.who.int/news-room/commentaries/detail/criteria-for-releasing-covid-19-patients-from-isolation

74. Direção Geral da Saúde. Plano nacional de preparação e resposta à doença por novo coronavírus (COVID-19). 2020.

75. Mehraeen E, Salehi MA, Behnezhad F, Moghaddam HR, SeyedAlinaghi S. Transmission Modes of COVID-19: A Systematic Review. Infect Disord Drug Targets. 15 de Novembro de 2020;

76. World Health Organisation. Transmission of SARS-CoV-2: implications for infection prevention precautions. 2020 Jul.

77. Meyerowitz EA, Richterman A, Gandhi RT, Sax PE. Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors. Ann Intern Med. Janeiro de 2021;174(1):69–79.

78. Coronavirus disease (COVID-19): How is it transmitted? [Internet]. [citado 25 de Maio

de 2021]. Disponível em: https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted

79. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect Dis. Abril de 2020;20(4):411–2.

80. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA [Internet]. 11 de Março de 2020 [citado 10 de Maio de 2021]; Disponível em: https://jamanetwork.com/journals/jama/fullarticle/2762997

81. Wu Y, Guo C, Tang L, Hong Z, Zhou J, Dong X, et al. Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol Hepatol. Maio de 2020;5(5):434–5.

82. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q, et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study. BMJ. 21 de Abril de 2020;369:m1443.

83. Sit THC, Brackman CJ, Ip SM, Tam KWS, Law PYT, To EMW, et al. Infection of dogs with SARS-CoV-2. Nature. Outubro de 2020;586(7831):776–8.

84. Newman A, Smith D, Ghai RR, Wallace RM, Torchetti MK, Loiacono C, et al. First Reported Cases of SARS-CoV-2 Infection in Companion Animals - New York, March-April 2020. MMWR Morb Mortal Wkly Rep. 12 de Junho de 2020;69(23):710–3.

85. Oreshkova N, Molenaar R-J, Vreman S, Harders F, Munnink BBO, Hakze R, et al. SARS-CoV2 infection in farmed mink, Netherlands, April 2020. bioRxiv. 18 de Maio de 2020;2020.05.18.101493.

86. Sooksawasdi Na Ayudhya S, Kuiken T. Reverse Zoonosis of COVID-19: Lessons From the 2009 Influenza Pandemic. Vet Pathol. Março de 2021;58(2):234–42.

87. European Centre for Disease Prevention and Control. Detection of new SARS-CoV-2 variants related to mink. 2020 Nov p. 19. (Rapid Risk Assessment).

88. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science. 8 de Janeiro de 2021;371(6525):172–7.

89. Marzoli F, Bortolami A, Pezzuto A, Mazzetto E, Piro R, Terregino C, et al. A systematic review of human coronaviruses survival on environmental surfaces. Sci Total Environ. 3 de Março de 2021;778:146191.

90. Bedrosian N, Mitchell E, Rohm E, Rothe M, Kelly C, String G, et al. A Systematic Review of Surface Contamination, Stability, and Disinfection Data on SARS-CoV-2 (Through July 10, 2020). Environ Sci Technol. 6 de Abril de 2021;55(7):4162–73.

91. Yanes-Lane M, Winters N, Fregonese F, Bastos M, Perlman-Arrow S, Campbell JR, et al. Proportion of asymptomatic infection among COVID-19 positive persons and their transmission potential: A systematic review and meta-analysis. PLoS One. 2020;15(11):e0241536.

92. Health Information ans Quality Authority. Evidence summary for the incubation period of COVID-19, or time to first positive test, in individuals exposed to SARS-CoV-2. 2020 Nov.

93. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. Maio de 2020;26(5):672–5.

94. Fontana LM, Villamagna AH, Sikka MK, McGregor JC. Understanding viral shedding of severe acute respiratory coronavirus virus 2 (SARS-CoV-2): Review of current literature. Infect Control Hosp Epidemiol. Junho de 2021;42(6):659–68.

95. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. Maio de 2020;581(7809):465–9.

96. Liu Y, Yan L-M, Wan L, Xiang T-X, Le A, Liu J-M, et al. Viral dynamics in mild and severe cases of COVID-19. Lancet Infect Dis. Junho de 2020;20(6):656–7.

97. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 28 de Março de 2020;395(10229):1054–62.

98. Xu K, Chen Y, Yuan J, Yi P, Ding C, Wu W, et al. Factors Associated With Prolonged Viral RNA Shedding in Patients with Coronavirus Disease 2019 (COVID-19). Clin Infect Dis. 28 de Julho de 2020;71(15):799–806.

99. Van Vinh Chau N, Lam VT, Dung NT, Yen LM, Minh NNQ, Hung LM, et al. The Natural History and Transmission Potential of Asymptomatic Severe Acute Respiratory Syndrome Coronavirus 2 Infection. Clin Infect Dis. 17 de Dezembro de 2020;71(10):2679–87.

100. Xiao AT, Tong YX, Zhang S. Profile of RT-PCR for SARS-CoV-2: A Preliminary Study From 56 COVID-19 Patients. Clin Infect Dis. 19 de Novembro de 2020;71(16):2249–51.

101. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. Lancet Microbe. Janeiro de 2021;2(1):e13–22.

102. Borges V, Isidro J, Cunha M, Cochicho D, Martins L, Banha L, et al. Long-Term Evolution of SARS-CoV-2 in an Immunocompromised Patient with Non-Hodgkin Lymphoma. mSphere. 6(4):e00244-21.

103. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. New England Journal of Medicine [Internet]. 11 de Novembro de 2020 [citado 16 de Novembro de 2021]; Disponível em: https://www.nejm.org/doi/10.1056/NEJMc2031364

104. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. Abril de 2021;592(7853):277–82.

105. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 15 de Janeiro de 2021;371(6526):eabe2424.

106. Laxminarayan R, Wahl B, Dudala SR, Gopal K, B CM, Neelima S, et al. Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science. 6 de Novembro de 2020;370(6517):691–7.

107. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nat Med. Novembro de 2020;26(11):1714–9.

108. Koh WC, Naing L, Chaw L, Rosledzana MA, Alikhan MF, Jamaludin SA, et al. What

do we know about SARS-CoV-2 transmission? A systematic review and meta-analysis of the secondary attack rate and associated risk factors. PLoS One. 8 de Outubro de 2020;15(10):e0240205.

109. Pung R, Chiew CJ, Young BE, Chin S, Chen MI-C, Clapham HE, et al. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. Lancet. 28 de Março de 2020;395(10229):1039–46.

110. Hall VJ, Foulkes S, Charlett A, Atti A, Monk EJM, Simmons R, et al. SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). Lancet. 17 de Abril de 2021;397(10283):1459–69.

111. Hansen CH, Michlmayr D, Gubbels SM, Mølbak K, Ethelberg S. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. Lancet. 27 de Março de 2021;397(10280):1204–12.

112. Harrington D, Kele B, Pereira S, Couto-Parada X, Riddell A, Forbes S, et al. Confirmed Reinfection With Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variant VOC-202012/01. Clin Infect Dis. 9 de Janeiro de 2021;ciab014.

113. Zucman N, Uhel F, Descamps D, Roux D, Ricard J-D. Severe reinfection with South African SARS-CoV-2 variant 501Y.V2: A case report. Clin Infect Dis. 10 de Fevereiro de 2021;ciab129.

114. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. Março de 2020;579(7798):265–9.

115. Eurosurveillance | Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR [Internet]. [citado 23 de Janeiro de 2021]. Disponível em: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.3.2000045

116. Directorate-General of Health - DGS. Norm 004/2020. 2020.

117. Watson J, Whiting PF, Brush JE. Interpreting a covid-19 test result. BMJ. 12 de Maio

de 2020;m1808.

118. Directorate-General of Health - DGS. Pandemia da gripe. 2007.

119. Vaccines and treatment of COVID-19 [Internet]. European Centre for Disease Prevention and Control. [citado 18 de Fevereiro de 2021]. Disponível em: https://www.ecdc.europa.eu/en/covid-19/latest-evidence/vaccines-and-treatment

120. Wang M, Wu T, Zuo Z, You Y, Yang X, Pan L, et al. Evaluation of current medical approaches for COVID-19: a systematic review and meta-analysis. BMJ Support Palliat Care. Março de 2021;11(1):45–52.

121. Lan S-H, Lai C-C, Huang H-T, Chang S-P, Lu L-C, Hsueh P-R. Tocilizumab for severe COVID-19: a systematic review and meta-analysis. Int J Antimicrob Agents. Setembro de 2020;56(3):106103.

122. WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group, Sterne JAC, Murthy S, Diaz JV, Slutsky AS, Villar J, et al. Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19: A Meta-analysis. JAMA. 6 de Outubro de 2020;324(13):1330–41.

123. Singh AK, Singh A, Singh R, Misra A. Molnupiravir in COVID-19: A systematic review of literaturef. Diabetes Metab Syndr. 30 de Outubro de 2021;15(6):102329.

124. Direção Geral da Saúde. Norma nº 001/2021: Campanha de Vacinação contra a COVID-19 Vacina SPIKEVAX®. 2021.

125. Direção Geral da Saúde. Norma nº 003/2021: Campanha de Vacinação contra a COVID-19 Vacina VAXZEVRIA. 2021.

126. Direção Geral da Saúde. Norma nº 004/2021: Campanha de Vacinação contra a COVID-19 COVID-19 Vaccine Janssen®. 2021.

127. Direção Geral da Saúde. Norma nº 021/2021: Campanha de Vacinação contra a COVID-19 Vacina COMIRNATY®. 2020.

128. Real-World Evidence Confirms High Effectiveness of Pfizer-BioNTech COVID-19 Vaccine and Profound Public Health Impact of Vaccination One Year After Pandemic Declared | pfpfizeruscom [Internet]. [citado 25 de Maio de 2021]. Disponível em:

https://www.pfizer.com/news/press-release/press-release-detail/real-world-evidence-confirms-high-effectiveness-pfizer

129. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. New England Journal of Medicine. 4 de Fevereiro de 2021;384(5):403–16.

130. Voysey M, Clemens SAC, Madhi SA, Weckx LY, Folegatti PM, Aley PK, et al. Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: a pooled analysis of four randomised trials. The Lancet. 6 de Março de 2021;397(10277):881–91.

131. Harder T, Külper-Schiek W, Reda S, Treskova-Schwarzbach M, Koch J, Vygen-Bonnet S, et al. Effectiveness of COVID-19 vaccines against SARS-CoV-2 infection with the Delta (B.1.617.2) variant: second interim results of a living systematic review and meta-analysis, 1 January to 25 August 2021. Euro Surveill. Outubro de 2021;26(41).

132. CDC. Science Brief: COVID-19 Vaccines and Vaccination [Internet]. Centers for Disease Control and Prevention. 2020 [citado 9 de Junho de 2021]. Disponível em: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/fully-vaccinated-people.html

133. Shah ASV, Gribben C, Bishop J, Hanlon P, Caldwell D, Wood R, et al. Effect of vaccination on transmission of COVID-19: an observational study in healthcare workers and their households. medRxiv. 21 de Março de 2021;2021.03.11.21253275.

134. Harris RJ, Hall JA, Zaidi A, Andrews NJ, Dunbar JK. Impact of vaccination on household transmission of SARS-COV-2 in England. :32.

135. Levine-Tiefenbrun M, Yelin I, Katz R, Herzel E, Golan Z, Schreiber L, et al. Decreased SARS-CoV-2 viral load following vaccination [Internet]. Infectious Diseases (except HIV/AIDS); 2021 Fev [citado 9 de Junho de 2021]. Disponível em: http://medrxiv.org/lookup/doi/10.1101/2021.02.06.21251283

136. Jones NK, Rivett L, Seaman S, Samworth RJ, Warne B, Workman C, et al. Single-dose BNT162b2 vaccine protects against asymptomatic SARS-CoV-2 infection. van

der Meer JW, editor. eLife. 8 de Abril de 2021;10:e68808.

137. McEllistrem MC, Clancy CJ, Buehrle DJ, Lucas A, Decker BK. Single dose of a mRNA SARS-CoV-2 vaccine is associated with lower nasopharyngeal viral load among nursing home residents with asymptomatic COVID-19. Clin Infect Dis. 26 de Março de 2021;

138. Petter E, Mor O, Zuckerman N, Oz-Levi D, Younger A, Aran D, et al. Initial real world evidence for lower viral load of individuals who have been vaccinated by BNT162b2. medRxiv. 8 de Fevereiro de 2021;2021.02.08.21251329.

139. Levine-Tiefenbrun M, Yelin I, Katz R, Herzel E, Golan Z, Schreiber L, et al. Initial report of decreased SARS-CoV-2 viral load after inoculation with the BNT162b2 vaccine. Nat Med. Maio de 2021;27(5):790–2.

140. The effects of virus variants on COVID-19 vaccines [Internet]. [citado 9 de Junho de 2021]. Disponível em: https://www.who.int/news-room/feature-stories/detail/the-effects-of-virus-variants-on-covid-19-vaccines

141. Noori M, Nejadghaderi SA, Arshi S, Carson-Chahhoud K, Ansarin K, Kolahi A-A, et al. Potency of BNT162b2 and mRNA-1273 vaccine-induced neutralizing antibodies against severe acute respiratory syndrome-CoV-2 variants of concern: A systematic review of in vitro studies. Rev Med Virol. 19 de Julho de 2021;e2277.

142. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci U S A. 20 de Abril de 2004;101(16):6146–51.

143. European Centre for Disease Prevention and Control. Contact tracing for COVID-19: current evidence, options for scale-up and an assessment of resources needed. 2020 Abr p. 9.

144. Organization WH. Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief, 29 March 2020. 2020 [citado 23 de Janeiro de 2021]; Disponível em: https://apps.who.int/iris/handle/10665/331616

145. Data on country response measures to COVID-19 [Internet]. European Centre for Disease Prevention and Control. 2020 [citado 27 de Dezembro de 2020]. Disponível em: https://www.ecdc.europa.eu/en/publications-data/download-data-response-

measures-covid-19

146. Public health considerations while resuming international travel [Internet]. [citado 27 de Dezembro de 2020]. Disponível em: https://www.who.int/news-room/articles-detail/public-health-considerations-while-resuming-international-travel

147. Epidemiology Working Group for NCIP Epidemic Response, Chinese Center for Disease Control and Prevention. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. Zhonghua Liu Xing Bing Xue Za Zhi. 10 de Fevereiro de 2020;41(2):145–51.

148. Gordis Epidemiology - 6th Edition [Internet]. [citado 20 de Abril de 2021]. Disponível em: https://www.elsevier.com/books/gordis-epidemiology/celentano/978-0-323-55229-5

149. Directorate-General of Health - DGS. Despacho n.º 12513-B/2019. 2019.

150. Surveillance Case Definitions | NNDSS [Internet]. [citado 20 de Abril de 2021]. Disponível em: https://wwwn.cdc.gov/nndss/case-definitions.html

151. Directorate-General of Health - DGS. Norm 020/2020. 2020.

152. Direção Geral da Saúde. Orientação nº 002A/2020. 2020.

153. WHO EMRO | IHR core capacities | About | International Health Regulations [Internet]. [citado 27 de Janeiro de 2020]. Disponível em: http://www.emro.who.int/international-health-regulations/about/ihr-core-capacities.html

154. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003 [Internet]. [citado 13 de Maio de 2021]. Disponível em: https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003

155. Autoridade de Saúde [Internet]. [citado 13 de Maio de 2021]. Disponível em: https://www.dgs.pt/a-dgs/autoridades-de-saude.aspx

156. Directorate-General of Health - DGS. Despacho n.º 5855/2014.

157. Miranda Ferrão, Rita; Fronteira, Inês; Sá Machado, Rita; Duarte, Eduardo. Surveillance of malaria in mainland Portugal in 2016 after the introduction of an

electronic system: is it enough? European Journal of Public Health. 2019;

158. Conducting a Field Investigation | Epidemic Intelligence Service | CDC [Internet]. 2019 [citado 20 de Maio de 2021]. Disponível em: https://www.cdc.gov/eis/field-epi-manual/chapters/Field-Investigation.html

159. Silman AJ, Macfarlane GJ, Macfarlane T. Epidemiological studies [Internet]. Epidemiological Studies: A Practical Guide. Oxford University Press; [citado 24 de Abril de 2021]. Disponível em: https://oxfordmedicine.com/view/10.1093/med/9780198814726.001.0001/med-9780198814726-chapter-10

160. Public Health England. Communicable disease outbreak management: operational guidance [Internet]. 2014 [citado 19 de Outubro de 2021]. Disponível em: https://www.gov.uk/government/publications/communicable-disease-outbreak-management-operational-guidance

161. Garry M, Hope L, Zajac R, Verrall AJ, Robertson JM. Contact Tracing: A Memory Task With Consequences for Public Health. Perspect Psychol Sci. 1 de Janeiro de 2021;16(1):175–87.

162. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. Eur Respir J. Janeiro de 2013;41(1):140–56.

163. Girum T, Lentiro K, Geremew M, Migora B, Shewamare S. Global strategies and effectiveness for COVID-19 prevention through contact tracing, screening, quarantine, and isolation: a systematic review. Trop Med Health [Internet]. 23 de Novembro de 2020 [citado 12 de Maio de 2021];48. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7680824/

164. Regime da criação, estruturação e funcionamento dos agrupamentos de centros de saúde do Serviço Nacional de Saúde [Internet]. Diário da República Eletrónico. [citado 9 de Junho de 2021]. Disponível em: https://dre.pt/web/guest/legislacao-consolidada/-/lc/118376180/202104190522/73662511/diploma/indice?p_p_state=maximized

165. : :: DL n.º 135/2013, de 04 de Outubro [Internet]. [citado 9 de Junho de 2021]. Disponível em:

https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=2025&tabela=leis&so_miolo=

166. Lei de Bases da Saúde [Internet]. Diário da República Eletrónico. [citado 9 de Junho de 2021]. Disponível em: https://dre.pt/web/guest/legislacao-consolidada/-/lc/57483775/201905100030/57494543/diploma/indice

167. SPMS. Manual de utilizador Trace COVID-19. 2020.

168. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. Abril de 2003;422(6934):835–47.

169. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med. Outubro de 2007;9(10):665–74.

170. Auffray C, Caulfield T, Griffin JL, Khoury MJ, Lupski JR, Schwab M. From genomic medicine to precision medicine: highlights of 2015. Genome Medicine. 29 de Janeiro de 2016;8(1):12.

171. Caudle KE, Gammal RS, Whirl-Carrillo M, Hoffman JM, Relling MV, Klein TE. Evidence and resources to implement pharmacogenetic knowledge for precision medicine. Am J Health Syst Pharm. 1 de Dezembro de 2016;73(23):1977–85.

172. Djordjevic N, Boccia S, Ádány R. Editorial: Translation of Genomic Results Into Public Health Practice. Front Public Health [Internet]. 2018 [citado 7 de Abril de 2021];6. Disponível em: https://www.frontiersin.org/articles/10.3389/fpubh.2018.00156/full#B2

173. Khoury MJ, Evans JP. A Public Health Perspective on a National Precision Medicine Cohort: Balancing Long-term Knowledge Generation With Early Health Benefit. JAMA. 2 de Junho de 2015;313(21):2117–8.

174. Khoury MJ, Iademarco MF, Riley WT. Precision Public Health for the Era of Precision Medicine. Am J Prev Med. Março de 2016;50(3):398–401.

175. Boccia S, Mc Kee M, Adany R, Boffetta P, Burton H, Cambon-Thomsen A, et al. Beyond public health genomics: proposals from an international working group. European Journal of Public Health. 1 de Dezembro de 2014;24(6):877–9.

176. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, et al. From public health genomics to precision public health: a 20-year journey. Genet Med. Junho de 2018;20(6):574–82.

177. says DST. Precision Public Health: What Is It? | | Blogs | CDC [Internet]. [citado 25 de Maio de 2021]. Disponível em: https://blogs.cdc.gov/genomics/2018/05/15/precision-public-health-2/

178. Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management | medRxiv [Internet]. [citado 7 de Abril de 2021]. Disponível em: https://www.medrxiv.org/content/10.1101/2020.09.28.20201475v2

179. Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science [Internet]. 5 de Fevereiro de 2021 [citado 7 de Abril de 2021];371(6529). Disponível em: https://science.sciencemag.org/content/371/6529/eabe3261

180. Popovich KJ, Snitkin ES. Whole Genome Sequencing—Implications for Infection Prevention and Outbreak Investigations. Curr Infect Dis Rep. 9 de Março de 2017;19(4):15.

181. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen Genomics in Public Health. New England Journal of Medicine [Internet]. 25 de Dezembro de 2019 [citado 16 de Novembro de 2021]; Disponível em: https://www.nejm.org/doi/full/10.1056/NEJMsr1813907

182. Advanced Molecular Detection (AMD) and Response to Infectious Disease Outbreaks [Internet]. 2021 [citado 7 de Abril de 2021]. Disponível em: https://www.cdc.gov/amd/index.html

183. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP. INSaFLU: an automated open web-based bioinformatics suite “from-reads” for influenza whole-genome-sequencing-based surveillance. Genome Med. Dezembro de 2018;10(1):46.

184. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time

nanopore sequencing analysis. Genome Med. 29 de Setembro de 2015;7:99.

185. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. Fevereiro de 2016;530(7589):228–32.

186. Di Paola N, Sanchez-Lockhart M, Zeng X, Kuhn JH, Palacios G. Viral genomics in Ebola virus research. Nature Reviews Microbiology. Julho de 2020;18(7):365–78.

187. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, et al. Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli. Journal of Clinical Microbiology. 1 de Maio de 2014;52(5):1501–10.

188. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, et al. Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing Escherichia coli Other than Serogroup O157. Front Microbiol. 2016;7:258.

189. Berenger BM, Berry C, Peterson T, Fach P, Delannoy S, Li V, et al. The utility of multiple molecular methods including whole genome sequencing as tools to differentiate Escherichia coli O157:H7 outbreaks. Euro Surveill. 2015;20(47).

190. Graham RMA, Doyle CJ, Jennison AV. Real-time investigation of a Legionella pneumophila outbreak using whole genome sequencing. Epidemiol Infect. Novembro de 2014;142(11):2347–51.

191. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biology. 30 de Maio de 2015;16(1):114.

192. Multi-country outbreak of Salmonella Enteritidis phage type 8, MLVA type 2-9-7-3-2 and 2-9-6-3-2 infections [Internet]. Autorità europea per la sicurezza alimentare. [citado 26 de Abril de 2021]. Disponível em: https://www.efsa.europa.eu/it/supporting/pub/en-1110

193. Crowe SJ, Green A, Hernandez K, Peralta V, Bottichio L, Defibaugh-Chavez S, et al. Utility of Combining Whole Genome Sequencing with Traditional Investigational Methods To Solve Foodborne Outbreaks of Salmonella Infections Associated with Chicken: A New Tool for Tackling This Challenging Food Vehicle. J Food Prot. Abril

de 2017;80(4):654–60.

194. Cartwright EJ, Jackson KA, Johnson SD, Graves LM, Silk BJ, Mahon BE. Listeriosis outbreaks and associated food vehicles, United States, 1998-2008. Emerg Infect Dis. Janeiro de 2013;19(1):1–9; quiz 184.

195. Price JR, Golubchik T, Cole K, Wilson DJ, Crook DW, Thwaites GE, et al. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of Staphylococcus aureus in an intensive care unit. Clin Infect Dis. Março de 2014;58(5):609–18.

196. Guthrie JL, Strudwick L, Roberts B, Allen M, McFadzen J, Roth D, et al. Whole genome sequencing for improved understanding of Mycobacterium tuberculosis transmission in a remote circumpolar region. Epidemiol Infect. Janeiro de 2019;147:e188.

197. Althomsons SP, Hill AN, Harrist AV, France AM, Powell KM, Posey JE, et al. Statistical Method to Detect Tuberculosis Outbreaks among Endemic Clusters in a Low-Incidence Setting. Emerg Infect Dis. Março de 2018;24(3):573–5.

198. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 24 de Fevereiro de 2011;364(8):730–9.

199. Guthrie JL, Delli Pizzi A, Roth D, Kong C, Jorgensen D, Rodrigues M, et al. Genotyping and Whole-Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia, Canada, 2005-2014. J Infect Dis. 24 de Agosto de 2018;218(7):1155–63.

200. Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. PLoS One. 2018;13(4):e0195413.

201. Correia CR, Marçal M, Vieira F, Santos E, Novais C, Maria AT, et al. Congenital SARS-CoV-2 Infection in a Neonate With Severe Acute Respiratory Syndrome. The Pediatric Infectious Disease Journal. Dezembro de 2020;39(12):e439.

202. Macedo R, Pinto M, Borges V, Nunes A, Oliveira O, Portugal I, et al. Evaluation of a

gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant Mycobacterium tuberculosis. Tuberculosis. 1 de Março de 2019;115:81–8.

203. Borges V, Nunes A, Sampaio DA, Vieira L, Machado J, Simões MJ, et al. Legionella pneumophila strain associated with the first evidence of person-to-person transmission of Legionnaires' disease: a unique mosaic genetic backbone. Sci Rep. 19 de Maio de 2016;6:26261.

204. Pinto M, Borges V, Isidro J, Rodrigues JC, Vieira L, Borrego MJ, et al. Neisseria gonorrhoeae clustering to reveal major European whole-genome-sequencing-based genogroups in association with antimicrobial resistance. Microbial Genomics. 7(2):000481.

205. Borges V, Isidro J, Correia C, Cordeiro D, Vieira L, Lodhia Z, et al. Transcontinental Dissemination of the L2b/D-Da Recombinant Chlamydia trachomatis Lymphogranuloma venereum (LGV) Strain: Need of Broad Multi-Country Molecular Surveillance. Clinical Infectious Diseases. 15 de Agosto de 2021;73(4):e1004–7.

206. Silveira L, Pinto M, Isidro J, Pista Â, Themudo P, Vieira L, et al. Multidrug-Resistant Salmonella enterica Serovar Rissen Clusters Detected in Azores Archipelago, Portugal. International Journal of Genomics. 16 de Dezembro de 2019;2019:e1860275.

207. Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. Evol Med Public Health. 9 de Junho de 2014;2014(1):96–108.

208. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. Whole-genome sequencing in outbreak analysis. Clin Microbiol Rev. Julho de 2015;28(3):541–63.

209. Borges V, Isidro J, Cortes-Martins H, Duarte S, Vieira L, Leite R, et al. Massive dissemination of a SARS-CoV-2 Spike Y839 variant in Portugal. Emerging Microbes & Infections. 1 de Janeiro de 2020;9(1):2488–96.

210. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. Setembro de 2020;26(9):1405–10.

211. National Academies of Sciences, Engineering, and Medicine; Division on Earth and Life Studies; Board on Life Sciences; Health and Medicine Division; Board on Health Sciences Policy; Committee on Data Needs to Monitor Evolution of SARS-CoV-2. Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies [Internet]. Washington (DC): National Academies Press (US); 2020 [citado 20 de Maio de 2021]. Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK562770/

212. Aggarwal D, Myers R, Hamilton WL, Bharucha T, Tumelty NM, Brown CS, et al. The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. The Lancet Microbe [Internet]. 29 de Setembro de 2021 [citado 1 de Outubro de 2021];0(0). Disponível em: https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(21)00208-1/abstract

213. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria NR, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 31 de Julho de 2020;369(6503):582–7.

214. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. Nat Microbiol. Julho de 2020;5(7):876–7.

215. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med. 9 de Dezembro de 2020;12(573).

216. Snell LB, Fisher CL, Taj U, Stirrup O, Merrick B, Alcolea-Medina A, et al. Combined epidemiological and genomic analysis of nosocomial SARS-CoV-2 infection early in the pandemic and the role of unidentified cases in transmission. Clin Microbiol Infect [Internet]. 13 de Agosto de 2021 [citado 29 de Setembro de 2021]; Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8361005/

217. Page AJ, Mather AE, Le-Viet T, Meader EJ, Alikhan N-F, Kay GL, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. Microb Genom. Junho de 2021;7(6).

218. Pedro N, Fernandes V, Cavadas B, Guimarães JT, Barros H, Tavares M, et al. Field

and Molecular Epidemiology: How Viral Sequencing Changed Transmission Inferences in the First Portuguese SARS-CoV-2 Infection Cluster. Viruses. Junho de 2021;13(6):1116.

219. Borges V, Isidro J, Macedo F, Neves J, Silva L, Paiva M, et al. Nosocomial Outbreak of SARS-CoV-2 in a “Non-COVID-19” Hospital Ward: Virus Genome Sequencing as a Key Tool to Understand Cryptic Transmission. Viruses. Abril de 2021;13(4):604.

220. Correia AM, Borges V, Isidro J, Lima AR, Fernandes A, Godinho ML, et al. Potential recurrence of COVID-19 in a healthcare professional: SARS-CoV-2 genome sequencing confirms contagiousness after re-positivity. International Journal of Infectious Diseases. 1 de Novembro de 2021;112:318–20.

221. Borges V, Sousa C, Menezes L, Gonçalves AM, Picão M, Almeida JP, et al. Tracking SARS-CoV-2 lineage B.1.1.7 dissemination: insights from nationwide spike gene target failure (SGTF) and spike gene late detection (SGTL) data, Portugal, week 49 2020 to week 3 2021. Eurosurveillance. 11 de Março de 2021;26(10):2100131.

222. Quick J. nCoV-2019 sequencing protocol. 22 de Janeiro de 2020 [citado 21 de Maio de 2021]; Disponível em: https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w

223. Nextclade [Internet]. [citado 14 de Novembro de 2021]. Disponível em: https://clades.nextstrain.org

224. O’Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. Virus Evolution. 5 de Julho de 2021;veab064.

225. Campbell F, Jombart T, Randhawa N, Sudre B, Nagraj VP, Crellen T, et al. epicontacts: Handling, Visualisation and Analysis of Epidemiological Contacts [Internet]. 2021 [citado 31 de Outubro de 2021]. Disponível em: https://CRAN.R-project.org/package=epicontacts

226. Szklo M, Javier Nieto F. Epidemiology: beyond the basics. 2nd Ed. Sudbury: Jones and Bartlett Publishers; 2007.

227. Naike Wang. How to Conduct a Meta-Analysis of Proportions in R: A Comprehensive Tutorial [Internet]. 2018 [citado 29 de Junho de 2021]. Disponível em:

https://www.researchgate.net/publication/325486099_How_to_Conduct_a_Meta-Analysis_of_Proportions_in_R_A_Comprehensive_Tutorial

228. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. Res Synth Methods. Setembro de 2019;10(3):476–83.

229. Vitor Borges, Joana Isidro, Helena Cortes-Martins, Luis Vieira, Ricardo Leite, Isabel Gordo, et al. On the track of the D839Y mutation in the SARS-CoV-2 Spike fusion peptide: emergence and geotemporal spread of a highly prevalent variant in Portugal | medRxiv. [citado 1 de Outubro de 2020]; Disponível em: https://www.medrxiv.org/content/10.1101/2020.08.10.20171884v1

230. Borges V, Isidro J, Trovão NS, Duarte S, Cortes-Martins H, Martiniano H, et al. The early dynamics of the SARS-CoV-2 epidemic in Portugal. medRxiv. 23 de Fevereiro de 2021;2021.02.22.21252216.

231. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance. 13 de Agosto de 2020;25(32):2001410.

232. Lehnertz NB, Wang X, Garfin J, Taylor J, Zipprich J, VonBank B, et al. Transmission Dynamics of Severe Acute Respiratory Syndrome Coronavirus 2 in High-Density Settings, Minnesota, USA, March–June 2020. Emerg Infect Dis. Agosto de 2021;27(8):2052–63.

233. Stirrup O, Hughes J, Parker M, Partridge DG, Shepherd JG, Blackstone J, et al. Rapid feedback on hospital onset SARS-CoV-2 infections combining epidemiological and sequencing data. Elife. 29 de Junho de 2021;10:e65828.

234. Loconsole D, Sallustio A, Accogli M, Leaci A, Sanguedolce A, Parisi A, et al. Investigation of an outbreak of symptomatic SARS-CoV-2 VOC 202012/01-lineage B.1.1.7 infection in healthcare workers, Italy. Clin Microbiol Infect. Agosto de 2021;27(8):1174.e1-1174.e4.

235. Cheng VC-C, Fung KS-C, Siu GK-H, Wong S-C, Cheng LS-K, Wong M-S, et al. Nosocomial Outbreak of Coronavirus Disease 2019 by Possible Airborne

Transmission Leading to a Superspreading Event. Clin Infect Dis. 15 de Setembro de 2021;73(6):e1356–64.

236. Lumley SF, Constantinides B, Sanderson N, Rodger G, Street TL, Swann J, et al. Epidemiological data and genome sequencing reveals that nosocomial transmission of SARS-CoV-2 is underestimated and mostly mediated by a small number of highly infectious individuals. J Infect. Outubro de 2021;83(4):473–82.

237. Wong RCW, Lee MKP, Siu GKH, Lee LK, Leung JSL, Leung ECM, et al. Healthcare workers acquired COVID-19 disease from patients? An investigation by phylogenomics. J Hosp Infect. Setembro de 2021;115:59–63.

238. Francis RV, Billam H, Clarke M, Yates C, Tsoleridis T, Berry L, et al. The impact of real-time whole genome sequencing in controlling healthcare-associated SARS-CoV-2 outbreaks. J Infect Dis. 23 de Setembro de 2021;jiab483.

239. van Hensbergen M, den Heijer CDJ, Wolffs P, Hackert V, ter Waarbeek HLG, Oude Munnink BB, et al. COVID-19: first long-term care facility outbreak in the Netherlands following cross-border introduction from Germany, March 2020. BMC Infect Dis. 4 de Maio de 2021;21:418.

240. Taylor J. Serial Testing for SARS-CoV-2 and Virus Whole Genome Sequencing Inform Infection Risk at Two Skilled Nursing Facilities with COVID-19 Outbreaks — Minnesota, April–June 2020. MMWR Morb Mortal Wkly Rep [Internet]. 2020 [citado 7 de Abril de 2021];69. Disponível em: https://www.cdc.gov/mmwr/volumes/69/wr/mm6937a3.htm

241. Ladhani SN, Chow JY, Janarthanan R, Fok J, Crawley-Boevey E, Vusirikala A, et al. Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care Homes. J Infect. Outubro de 2020;81(4):621–4.

242. Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science. 5 de Fevereiro de 2021;371(6529).

243. Pattabiraman C, Habib F, P. K. H, Rasheed R, Prasad P, Reddy V, et al. Genomic epidemiology reveals multiple introductions and spread of SARS-CoV-2 in the Indian

state of Karnataka. PLoS One. 17 de Dezembro de 2020;15(12):e0243412.

244. Weissberg D, Böni J, Rampini SK, Kufner V, Zaheri M, Schreiber PW, et al. Does respiratory co-infection facilitate dispersal of SARS-CoV-2? investigation of a super-spreading event in an open-space office. Antimicrob Resist Infect Control. 2 de Dezembro de 2020;9(1):191.

245. Boogaard LH, Sikkema RS, van Beek JHGM, Brockhoff HJ, Dalebout E, de Heus B, et al. A mixed-methods approach to elucidate SARS-CoV-2 transmission routes and clustering in outbreaks in native workers and labour migrants in the fruit and vegetable packaging industry in South Holland, the Netherlands, May to July 2020. Int J Infect Dis. Agosto de 2021;109:24–32.

246. National Institute of Health (INSA) Dr. Ricardo Jorge. Genetic diversity of the novel coronavirus SARS-CoV-2 (COVID-19) in Portugal [Internet]. [citado 29 de Julho de 2021]. Disponível em: https://insaflu.insa.pt/covid19/

247. European Centre for Disease Prevention and Control. Expert opinion on whole genome sequencing for public health surveillance: strategy to harness whole genome sequencing to strengthen EU outbreak investigations and public health surveillance. [Internet]. LU: Publications Office; 2016. Disponível em: doi 10.2900/12442

248. Lane CR, Sherry NL, Porter AF, Duchene S, Horan K, Andersson P, et al. Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study. Lancet Public Health. Agosto de 2021;6(8):e547–56.

249. Lemey P, Ruktanonchai N, Hong SL, Colizza V, Poletto C, Van den Broeck F, et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature. Julho de 2021;595(7869):713–7.

250. Salvatore PP, Dawson P, Wadhwa A, Rabold EM, Buono S, Dietrich EA, et al. Epidemiological Correlates of Polymerase Chain Reaction Cycle Threshold Values in the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Clin Infect Dis. 1 de Junho de 2021;72(11):e761–7.

251. Rousseau MJ, Evans JC. Key statistical assumptions and methods in one-arm meta-analyses with binary endpoints and low event rates, including a real-life example in

the area of endoscopic colonic stenting. Schumacher U, editor. null. 1 de Janeiro de 2017;4(1):1334318.

252. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol. Setembro de 2017;13(9):e1005697.

253. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 25 de Março de 2008;5(3):e74.

8. Appendices

Appendix 1 Variable's operationalization plan

Table 1.1. Variables' operationalization plan

Variable name	Designation	Values	Type
Individual Unit of Analysis			
ID	Identification	1,2,3…	Numerical
Date_EE	Date of introduction of the epidemiological investigation data	dd-mm-yyyy	date
NHS	National Health System Number	xxxxxxxxx	Numerical
Birth_date	Date of birth of the case	dd-mm-yyyy	date
Healthcare_prof	Is the patient a healthcare professional?	Yes (1)	Categorical
		No (0)	
Healthcare_prof_cat	What is the healthcare professional category?	Operational assistant	Categorical
		Medical Doctor	
		Nurse	
		Technical assistant	
		Pharmaceutical	
		Dentist	
		Senior Health Technician	
		Firefighter (pre hospital healthcare providers),	
		Not Available (NA)	
Healthcare_institution	Which healthcare institution class does he/she works for?	Hospital	Categorical
		Public Health Unit	
		Private clinic	
		Firefighters	
		Social care	
		Pharmacy	
Nursing_home	Is the patient a resident or worker in the nursing home?	Resident	Categorical
		Professional	
		No	
		NA	
School	Is the patient a student or professional in a school?	Professional	Categorical
		Student	
		No	
		NA	
Postal_code		XXXX-XXX	Numerical
Age	Age of the individual (years)	0-150	Numerical
Sex	Gender of the individual	Male (0)	Categorical
		Female (1)	
Municipality	Municipality where the individual is isolated	Águeda (1)	Categorical
		Anadia (2)	
		Albergaria a Velha (3)	
		Aveiro (4)	
		Estarreja (5)	
		Ílhavo (6)	
		Murtosa (7)	
		Oliveira do Bairro (8)	

		Ovar (9) Sever do Vouga (10) Vagos (11)	
Symptoms_onset	Date of the symptoms onset?	dd-mm-yyyy	date
Asymptomatic		Yes (1)	Categorical
		No (0)	
Fever		Yes (1)	Categorical
		No (0)	
Non-productive cough		Yes (1)	Categorical
		No (0)	
Productive cough		Yes (1)	Categorical
		No (0)	
Odynophagia		Yes (1)	Categorical
		No (0)	
Myalgia		Yes (1)	Categorical
		No (0)	
Dyspnoea		Yes (1)	Categorical
		No (0)	
Fatigue		Yes (1)	Categorical
		No (0)	
Headaches		Yes (1)	Categorical
		No (0)	
Thoracalgia		Yes (1)	Categorical
		No (0)	
Rhinorrhoea		Yes (1)	Categorical
		No (0)	
Anosmia		Yes (1)	Categorical
		No (0)	
Dysgeusia		Yes (1)	Categorical
		No (0)	
Diarrhoea		Yes (1)	Categorical
		No (0)	
Nausea		Yes (1)	Categorical
		No (0)	
Vomits		Yes (1)	Categorical
		No (0)	
Abdominal pain		Yes (1)	Categorical
		No (0)	
Arthralgia		Yes (1)	Categorical

		No (0)	
Conjunctivitis		Yes (1)	Categoric al
		No (0)	
Amnesia		Yes (1)	Categoric al
		No (0)	
Cutaneous disturbances		Yes (1)	Categoric al
		No (0)	
Other		Yes (1)	Categoric al
		No (0)	
Incub_period	Incubation period (median, IQR)	0-14	Numerical
	Number of days between the probable date of contagion and the onset of symtpms OR test		
Arterial_hypertension		Yes (1)	Categoric al
		No (0)	
Asthma		Yes (1)	Categoric al
		No (0)	
Diabetes		Yes (1)	Categoric al
		No (0)	
Dyslipidaemia		Yes (1)	Categoric al
		No (0)	
Cardiovascular_diseases		Yes (1)	Categoric al
		No (0)	
Neurological_disease		Yes (1)	Categoric al
		No (0)	
Chronic_kidney_disease		Yes (1)	Categoric al
		No (0)	
Chronic_haematological_ disease		Yes (1)	Categoric al
		No (0)	
Chronic_Obstructive_Pul monary Disease		Yes (1)	Categoric al
		No (0)	
Pregnancy		Yes (1)	Categoric al
		No (0)	
Malignancy		Yes (1)	Categoric al
		No (0)	
HIV		Yes (1)	Categoric al
		No (0)	
Other		Yes (1)	Categoric al

		No (0)	
Death	Did the patient died from COVID?	Yes	Categoric al
		No	
Date_death	What was the date of the patient death?	dd-mm-yyyy	date
Travelling		Yes (1)	Categoric al
		No (0)	
Contact_with_known_cas e		Yes (1)	Categoric al
		No (0)	
Unknown_source		Yes (1)	Categoric al
		No (0)	
Ascendant_ID	ID of the ascendant case (infector)	xxxxxxxxx	Categoric al
Conection	Reported connection between the case and its infector	Direct (1)	Categoric al
		Indirect (0)	
		Unknown (NA)	
Epi_chain	Epidemiologically linked chain of transmission attributed to the individual	A, B, C, D, E …	Categoric al
Gen_cluster	Genomical cluster attributed to the individual	1,2,3,4,5…	Categoric al
Cohabitant	If the epidemiological link (edge) presented is a cohabitant	Yes (1)	Categoric al
		No (0)	
Non-cohabitant	If the epidemiological link (edge) presented is a non-cohabitant acquaintance	Yes (1)	Categoric al
		No (0)	
Work_settings	If the epidemiological link (edge) presented is work associated	Yes (1)	Categoric al
		No (0)	
Healthcare	If the epidemiological link (edge) presented is healthcare associated	Yes (1)	Categoric al
		No (0)	
Nursinghome_School	Nursing_homes/Schoo l/Other social care institutions.	Yes (1)	Categoric al
		No (0)	
Travelling	If the epidemiological link (edge) presented is travel associated	Yes (1)	Categoric al
		No (0)	
Unknown	If the epidemiological link is not known	Yes (1)	Categoric al
		No (0)	

Direct	Indirect epidemiological link	Yes (1)	Categoric al
		No (0)	
Indirect	Indirect epidemiological link	Yes (1)	Categoric al
		No (0)	
Epiclusters Unit of Analysis			
Size_epichain	Number individuals present in every transmission chain	1-XX	Numerical
No_epichain	Number of epiclusters identified	1-XX	Numerical
No_epichain	Number of epiclusters with any genomic sequencing data	1-XX	Numerical
No_epichain_ inco	Number of incoherencies in epiclusters	0-XX	Numerical
No_gencluster	Number of genomic profiles	1-XX	Numerical
No_gencluster_epichain	Number of individuals in the genomic profile that belong to single dominant epidemiologically linked chain of transmission.	0-XX	Numerical
No_gencluster_inco	Number of incoherent individuals in the genomic profile.	0-XX	Numerical
Direct connections	Number of cases generated by one common ascendant case that report a direct connection	0-XX	Numerical
Indirect connections	Number of cases generated by one common ascendant case that report an indirect connection	0-XX	Numerical

Appendix 2 Characteristics of the sample

Table 2.1. Demographic, clinical presentation, context of the contagion, epidemiological link type, sanitary cordon area and case settings data for Baixo Vouga COVID-19 sporadic and within-cluster cases until 30 June 2020.

	Sporadics (N=782) n (%)	Cases within an epicluster (N=1143) n (%)	p-value
Demographic			
Age (at the day of the diagnostic test) – median (IQR)	55.61 [39.41;72.78]	55.53 [37.87;79.14]	0.316
Sex (Male)	315 (40.28%)	400 (35.00%)	0.021
Clinical presentation			
Symptomatic (Yes)	528 (67.52%)	736 (64.39%)	0.171
Fever (Yes)	71 (9.08%)	116 (10.15%)	0.484
Dry cough (Yes)	94 (12.02%)	181 (15.84%)	0.022
Productive cough (Yes)	6 (0.77%)	25 (2.19%)	0.025
Odynophagia (Yes)	26 (3.32%)	57 (4.99%)	0.099
Myalgia (Yes)	42 (5.37%)	111 (9.71%)	0.001
Dyspnoea (Yes)	30 (3.84%)	46 (4.02%)	0.929
Fatigue (Yes)	47 (6.01%)	52 (4.55%)	0.187
Headaches (Yes)	33 (4.22%)	135 (11.81%)	<0.001
Thoracalgia (Yes)	12 (1.53%)	29 (2.54%)	0.182
Rhinorrhoea (Yes)	13 (1.66%)	43 (3.76%)	0.011
Anosmia (Yes)	34 (4.35%)	78 (6.82%)	0.029
Dysgeusia (Yes)	25 (3.20%)	67 (5.86%)	0.010
Diarrhoea (Yes)	16 (2.05%)	39 (3.41%)	0.104
Amnesia (Yes)	2 (0.26%)	1 (0.09%)	0.570
Cutaneous disturbances (Yes)	1 (0.13%)	2 (0.17%)	1.000
Medical Conditions			
Medical conditions (Yes)	99 (37.93%)	101 (28.86%)	0.023
Hypertension (Yes)	44 (5.63%)	56 (4.90%)	0.547
Asthma (Yes)	5 (0.64%)	6 (0.52%)	0.766
Diabetes (Yes)	20 (2.56%)	31 (2.71%)	0.950
Dislipidemia (Yes)	20 (2.56%)	28 (2.45%)	1.000
Cardiovascular diseases (Yes)	26 (3.32%)	12 (1.05%)	0.001
Neurological disease (Yes)	17 (2.17%)	4 (0.35%)	<0.001
Chronic kidney disease (Yes)	3 (0.38%)	4 (0.35%)	1.000
Chronic hematological disease (Yes)	3 (0.38%)	0 (0.00%)	0.067
Chronic Obstructive Pulmonary Disease (Yes)	5 (0.64%)	1 (0.09%)	0.044
Pregnancy (Yes)	2 (0.26%)	2 (0.17%)	1.000
Malignancy (Yes)	17 (2.17%)	10 (0.87%)	0.029
other (Yes)	14 (1.79%)	22 (1.92%)	0.966
Putative source of acquisition			
Cohabitant (Yes)	43 (5.50%)	176 (15.40%)	<0.001
Healthcare (Yes)	14 (1.79%)	19 (1.66%)	0.973
Non cohabitant acquaintances (Yes)	16 (2.05%)	16 (1.40%)	0.364

Other (Yes)	6 (0.77%)	4 (0.35%)	0.333
Unknown (Yes)	662 (84.65%)	839 (73.40%)	<0.001
Work settings (Yes)	41 (5.24%)	89 (7.79%)	0.036
Transmission chain analysis			.
Direct	0 (.%)	433 (43.78%)	
Indirect	0 (.%)	556 (56.22%)	
Sanitary cordon area (at the date of diagnosis)	199 (25.45%)	419 (36.66%)	<0.001
Nursing home			
Nursing home resident or professional (Yes)	114 (14.58%)	245 (21.43%)	<0.001
Healthcare professional (Yes)	56 (7.16%)	144 (12.60%)	<0.001

*Unless indicated otherwise. IQR – Interquartile Range

Appendix 3 Characteristics of all the epiclusters with genomic data and sensitivity of the epidemiological investigation using genomics.

Table 3.1. Characteristics of all the epiclusters with genomic data and sensitivity of the epidemiological investigation using genomics.

Epicluster with genomic data	Cases within the same genomic cluster (No)	Cases outside the genomic cluster (No)	Sensitivity	Epicluster size
A	17	6	73.91%	280
B	22	8	73.33%	101
C	8	0	100.00%	44
D	1	0	100.00%	39
E	6	0	100.00%	35
F	2	0	100.00%	15
G	4	0	100.00%	11
H	3	0	100.00%	10
I	1	0	100.00%	10
J	6	0	100.00%	9
K	2	0	100.00%	7
L	1	0	100.00%	6
M	2	0	100.00%	6
N	3	0	100.00%	6
O	2	0	100.00%	5
P	1	0	100.00%	4
Q	2	0	100.00%	3
R	1	0	100.00%	3
S	1	0	100.00%	3
T	2	0	100.00%	3
U	2	0	100.00%	3
V	1	0	100.00%	3
W	1	0	100.00%	2
X	1	0	100.00%	2
Y	1	0	100.00%	2
Z	1	0	100.00%	2
A1	1	0	100.00%	2
A2	1	0	100.00%	6
28	96	14	87.27%	622

Appendix 4 Characteristics of the study sample (cases with available WGS)

Table 4.1. Demographic, clinical presentation, medical conditions, context of the contagion, epidemiological link type, sanitary cordon area and case settings data for Baixo Vouga COVID-19 cases with and without WGS data, until 30 June 2020.

	Cases without WGS data N=1797	Cases with available WGS data N=128	p-value
Demographic			
Age (at the day of the diagnostic test) – mean (sd)	55.45 [38.15;75.86]	56.05 [43.44;77.21]	0.300
Sex (Male)	654 (36.39%)	61 (47.66%)	0.014
Clinical presentation			
Symptomatic (Yes)	1164 (64.77%)	100 (78.12%)	0.003
Fever (Yes)	171 (9.52%)	16 (12.50%)	0.344
Dry cough (Yes)	252 (14.02%)	23 (17.97%)	0.271
Productive cough (Yes)	30 (1.67%)	1 (0.78%)	0.718
Odynophagia (Yes)	76 (4.23%)	7 (5.47%)	0.659
Myalgia (Yes)	143 (7.96%)	10 (7.81%)	1
Dyspnoea (Yes)	64 (3.56%)	12 (9.38%)	0.002
Fatigue (Yes)	88 (4.90%)	11 (8.59%)	0.105
Headaches (Yes)	154 (8.57%)	14 (10.94%)	0.45
Thoracalgia (Yes)	34 (1.89%)	7 (5.47%)	0.016
Rhinorrhoea (Yes)	51 (2.84%)	5 (3.91%)	0.417
Anosmia (Yes)	101 (5.62%)	11 (8.59%)	0.233
Dysgeusia (Yes)	84 (4.67%)	8 (6.25%)	0.553
Diarrhoea (Yes)	46 (2.56%)	9 (7.03%)	0.009
Amnesia (Yes)	2 (0.11%)	1 (0.78%)	0.187
Cutaneous disturbances (Yes)	2 (0.11%)	1 (0.78%)	0.187
Medical Conditions			
Medical conditions (Yes)	198 (32.57%)	2 (66.67%)	0.251
Hypertension (Yes)	99 (5.51%)	1 (0.78%)	0.034
Asthma (Yes)	11 (0.61%)	0 (0.00%)	1
Diabetes (Yes)	49 (2.73%)	2 (1.56%)	0.577
Dislipidemia (Yes)	48 (2.67%)	0 (0.00%)	0.072
Cardiovascular diseases (Yes)	38 (2.11%)	0 (0.00%)	0.175
Neurological disease (Yes)	20 (1.11%)	1 (0.78%)	1
Chronic kidney disease (Yes)	7 (0.39%)	0 (0.00%)	1
Chronic hematological disease (Yes)	3 (0.17%)	0 (0.00%)	1
Chronic Obstructive Pulmonary Disease (Yes)	6 (0.33%)	0 (0.00%)	1
Pregnancy (Yes)	4 (0.22%)	0 (0.00%)	1
Malignancy (Yes)	26 (1.45%)	1 (0.78%)	1
other (Yes)	36 (2.00%)	0 (0.00%)	0.168
Context of the contagion			
Cohabitant (Yes)	218 (12.13%)	1 (0.78%)	<0.001

Healthcare (Yes)	33 (1.84%)	0 (0.00%)	0.165
Non cohabitant acquaintances (Yes)	32 (1.78%)	0 (0.00%)	0.268
Other (Yes)	10 (0.56%)	0 (0.00%)	1
Unknown (Yes)	1374 (76.46%)	127 (99.22%)	<0.001
Work settings (Yes)	130 (7.23%)	0 (0.00%)	0.003
Epidemiological link type			0.549
Direct	385 (43.40%)	48 (47.06%)	
Indirect	502 (56.60%)	54 (52.94%)	
Sanitary cordon area (at the date of diagnosis)	523 (29.10%)	95 (74.22%)	<0.001
Case settings			
Nursing home resident or professional (Yes)	347 (19.31%)	12 (9.38%)	0.008
Healthcare professional (Yes)	191 (10.63%)	9 (7.03%)	0.255
Epicluster			
Cases within an epicluster	1033 (57.48%)	110 (85.94%)	<0.001

*Unless indicated otherwise. IQR – Interquartile Range. WGS – Whole Genome Sequencing.