Mathematical modelling to study infectious diseases: from understanding to prediction

Author: Martí Català Sabaté
Supervisors: Dr. Sergio Alonso Muñoz and Dr. Clara Prats Soler

PhD program on Comutational and Applied Phusics
Departament of Physics
Research group on Computational Biology and Complex System
Thesis by compendium of publications

Universitat Politècnica de Catalunya
Barcelona, September 2021

Mathematical modelling to study infectious diseases: from understanding to prediction

Author:

Martí Català Sabaté

Supervisors:

Dr. Sergio Alonso Muñoz

Dr. Clara Prats Soler

Thesis by compendium of publications

PhD program on Comutational and Applied Phusics

Departament of Physics

Research group on Computational Biology and Complex System

Universitat Politècnica de Catalunya

Barcelona, September 2021

Abstract

Each year 10 million people die from communicable diseases. They are infectious diseases caused by agents transmitted between individuals. Nowadays, the two infectious diseases that have the greatest impact are tuberculosis (TB) and coronavirus disease 2019 (COVID-19). According to the World Health Organization, TB killed 40 million individuals in the last 20 years. The COVID-19 pandemic has had an overwhelming effect on human life. It has caused millions of deaths and conditioned people's life since January 2020.

Mathematical and computational models are powerful tools in science to better understand, predict, and condition the dynamics of a desired system. In this thesis, we present a compendium of five publications where mathematical modelling is used to better understand and predict the dynamics of TB and COVID-19 at different spatio-temporal scales.

Although TB is a disease identified many years ago, its natural history is not fully understood yet. The main objectives of this thesis in this area are related with the understanding of the factors and processes that facilitate the triggering of an active disease from a latent tuberculosis infection. We also aim at improving understanding of the human-TB coexistence for more than 70,000 years and some of the particularities that have facilitated this coexistence.

We have built several models of the pulmonary TB infection at different spatial scales. At the alveolus level, we have seen that the correct balance of the immune response determines the outcome of the infection. At the secondary lobe level, we identified the distance to pulmonary membranes as an important factor to determine final lesion size. At the lung level, we have reproduced a dynamic hypothesis that explains the generation of secondary granulomas after the bronchial dissemination of the infectious bacilli from a preceding lesion. We have assessed the importance of lesion merging as a driving force for the triggering of the active disease.

In addition, we have modelled human-TB coexistence in the Paleolithic and Neolithic ages, and determined that female protection against TB was crucial for the survival of the human species. In the Neolithic age, new "modern" lineages emerged, displacing "ancient" ones. Mathematical modelling yields results that explain why this emergence was not

possible in the Paleolithic age.

When the COVID-19 pandemic started, there was a lack of monitoring systems to help control and manage the pandemic. In this thesis, we focus on several aims related to the assessment of the real incidence during the first wave, as well as on the building and testing of a short-term prediction model.

We developed a methodology to estimate the real incidence of COVID-19 based on the estimated lethality and the reported death series. We applied this to several European countries, after analyzing possible bias due to differing age structures. As well, we proposed and calibrated an empirical model based on the Gompertz growth that allows for reliable short-term forecasting at the country level.

This thesis demonstrates how mathematical and computational models can be used to predict and better understand important characteristics of infectious diseases such as TB and COVID-19.

Keywords: Mathematical modelling, Tuberculosis, COVID-19, Infectious diseases, Computational lung, Epidemiology, Biophysics, Mathematical biology.

Resum

Cada any 10 milions de persones es moren a causa de malalties transmissibles. Són malalties infeccioses causades per agents que es transmeten entre els diferents individus. Actualment, la tuberculosi (TB) i la COVID-19 són les dues malalties infeccioses que tenen un gran impacte. Segons les estimacions de l'Organització Mundial de la Salut, la TB ha causat la mort de gairebé 40 milions d'individus en els últims 20 anys. La pandèmia de la covid-19 ha afectat enormement la manera de viure de la població mundial. Des del gener del 2020 ha causat milions de morts i ha condicionat les vides i el comportament de les persones.

Els models matemàtics i computacionals són una eina fonamental que, en ciència, es poden usar per entendre, predir i/o condicionar la dinàmica d'un sistema en concret. En aquesta tesi presentem un compendi de cinc articles on s'usen els models matemàtics per entendre i predir les dinàmiques de la tuberculosi i la COVID-19 en diferents escales espai-temporals.

Tot i que la TB és una malaltia que es va identificar fa molts anys, alguns detalls de la seva història natural encara són desconeguts. L'objectiu principal d'aquesta tesi en relació a la TB és entendre els factors i processos que faciliten el pas de la infecció latent cap a malaltia activa. També hem intentat millorar el coneixement i identificar les particularitats dels 70000 anys de coexistència entre els humans i la TB.

S'han creat diferents models de la infecció tuberculosa pulmonar a diferents escales espacials. Al nivell alveolar, hem identificat que el correcte balanç entre la resposta immune i la resposta inflamatòria condiciona el resultat de la infecció. A escala de lòbul secundari, hem vist que la distància entre la lesió i la membrana pulmonar és un factor important que determinarà la seva mida final. A escala del pulmó, s'ha reproduït la hipòtesi dinàmica que ens permet explicar la generació de noves lesions a partir de disseminació bronquial de les lesions inicials. S'ha identificat el procés de fusió de lesions com un dels processos més importants que fa aparèixer lesions més grans i acaba originant la malaltia activa.

A més, hem modelitzat la coexistència entre els humans i la TB en el Paleolític i el Neolític. S'ha identificat que la protecció femenina envers la TB va ser crucial per la supervivència de l'espècie humana. En el neolític, van aparèixer soques "modernes" que van desplaçar les "antigues". Amb

models matemàtics s'ha pogut observar perquè aquesta aparició no va ser possible en el paleolític.

Quan la pandèmia de la COVID-19 va començar, els sistemes de vigilància que havien de servir per controlar i monitoritzar la pandèmia eren inexistents o deficients. En aquesta tesi hem treballat principalment en dos aspectes per ajudar a la monitorització de la pandèmia: determinar la incidència real de la primera onada i crear un model de prediccions a curt termini.

Hem desenvolupat una metodologia per determinar la incidència real que va tenir la COVID-19 basada en la letalitat de la malaltia i les sèries temporals de defuncions. Aquesta metodologia s'ha pogut aplicar a diversos països europeus, tenint en compte els possibles biaixos, per exemple, les diferents piràmides de població. S'ha proposat i calibrat un model empíric bastant en l'equació de Gompertz que ens permet fer una predicció dels casos a curt termini a nivell de país.

Aquesta tesi demostra com els models computacionals i matemàtics poden ajudar a predir i entendre millor les característiques de les malalties infeccioses usant com a exemple la tuberculosi i la COVID-19.

Resumen

Cada año 10 millones de personas mueren a causa de enfermedades contagiosas. Son enfermedades infecciosas causadas por agentes que se transmiten entre los diferentes individuos. Hoy en día, la tuberculosis (TB) y el COVID-19 son las dos enfermedades infecciosas que tienen un mayor impacto mundial. Según las estimaciones de la Organización Mundial de la Salud, la TB ha causado la muerte de casi 40 millones de individuos en los últimos 20 años. La pandemia del COVID-19 ha cambiado por completo la forma de vivir de la población mundial. Desde enero de 2020 ha causado millones de muertos y ha condicionado las vidas y el comportamiento de las personas.

Los modelos matemáticos y computacionales son una herramienta muy potente que, en ciencia, pueden ayudar a entender, predecir y/o condicionar la dinámica de un sistema en concreto. En esta tesis presentamos un compendio de cinco artículos donde se usan los modelos matemáticos para entender y predecir las dinámicas de la tuberculosis y el COVID-19 en diferentes escalas espacio-temporales.

Aunque la TB es una enfermedad que se identificó hace muchos años, algunos detalles de su historia natural aún son desconocidos. El objetivo principal de esta tesis en relación a la TB es entender los factores y procesos que facilitan el paso desde una infección latente a enfermedad activa. También hemos intentado mejorar el conocimiento e identificar las particularidades de los 70000 años de coexistencia entre los humanos y la TB.

Hemos creado diferentes modelos sobre la infección tuberculosa pulmonar a diferentes escalas espaciales. A nivel alveolar, hemos identificado que el correcto balance entre la respuesta inmune y la respuesta inflamatoria es determinante para el resultado de la infección. A escala del lóbulo secundario hemos visto que la distancia entre la lesión y la membrana pulmonar es un factor importante que determinará el tamaño final de la lesión. A escala del pulmón, se ha reproducido la hipótesis dinámica que nos permite explicar la generación de nuevas lesiones a partir de diseminación bronquial de las lesiones iniciales. Se ha identificado el proceso de fusión de lesiones como uno de los procesos más importantes que hace aparecer lesiones más grandes y acaba originando la enfermedad activa.

Además, hemos modelizado la coexistencia entre los humanos y la TB en el Paleolítico y el Neolítico. Se ha identificado que la protección femenina ante la TB fue crucial para la supervivencia de la especie humana. En el Neolítico, aparecieron cepas "modernas"que desplazaron las "antiguas". Con modelos matemáticos se ha podido observar que esta aparición no era posible en el Paleolítico.

Al iniciarse la pandemia del COVID-19, los sistemas de vigilancia que debían servir para controlar y monitorizar la pandemia eran inexistentes o deficientes. En esta tesis hemos trabajado principalmente en dos aspectos para ayudar a la monitorización de la pandemia: determinar la incidencia real de la primera ola y crear un modelo de predicciones a corto plazo.

Hemos desarrollado una metodología para determinar la incidencia real que tuvo el COVID-19 basada en la letalidad de la enfermedad y las series temporales de defunciones. Esta metodología se ha podido aplicar en varios países europeos, teniendo en cuenta los posibles sesgos, por ejemplo, las diferentes pirámides de población. También se ha usado y calibrado un modelo empírico basado en la ecuación de Gompertz que nos permite hacer una predicción de los casos a corto plazo a nivel de país.

Esta tesis demuestra cómo los modelos computacionales y matemáticos pueden ayudar a predecir y entender mejor las características de las enfermedades infecciosas usando como ejemplo la tuberculosis y el COVID-19.

Acknowledgements

Vull donar les gràcies de tot cor a totes les persones que m'han envoltat i ajudat aquests quatre darrers anys, sense totes elles no hauria estat possible seguir endavant i finalitzar aquesta tesi.

Primer de tot, l'agraïment més especial als meus directors. Al **Sergio Alonso**, que sempre m'ha dedicat hores i paciència, el seu esperit crític i constructiu m'ha aportat nous punts de vista i a explorar sempre una mica més enllà. I a la **Clara Prats**, què dir-ne! ..., ha estat una segona mare per a mi; amb un tracte excel·lent m'ha fet créixer com a persona i com a científic; després de sis anys treballant plegats només en puc parlar meravelles, no trobo la manera d'agrair-li tot el que ha fet i està fent per a mi. Moltes gràcies a tots dos.

No vull oblidar-me del **Dani López**, qui va ser el meu director durant sis mesos i que he continuat tenint sempre al costat quan li he demanat, vull destacar la seva preocupació per una part del món sovint oblidada.

Un reconeixement especial a tots ells i a la resta de companys del **BIOCOM-SC** que m'han acollit com un integrant més del grup i m'han deixat participar plenament, no solament en la feina sinó també en les seves vides. Gràcies especialment a l'**Enric Álvarez** que ha tingut un paper rellevant formant part del "team covid".

A tots els companys que han anat passant pel despatx de la universitat ajudant-me i fent aquesta tesi més gratificant, moltes gràcies Miquel, David, Eduardo, Miquel B, Francesc, Thomas, Katerina, Gustavo, Pablo, Dafni, Yunxian, Edneide, Allisson, Nura, Núria...

I també com no, el meu agraïment als professors del **departament de Física de la UPC**, el Blas, el Quim i el David. Vull comentar que treballar de professor en el departament ha estat una experiència professional molt satisfactòria, ho he gaudit molt.

També he comptat amb el suport incondicional de la **Unitat de Tuberculosi Experimental** (UTE) de Can Ruti. Moltes gràcies Pere-Joan, Cristina, Paula, Lili, Albert, Marta, Isabel, Jordi, Ricard i companyia, sense ells la tesi no hauria estat possible. En especial vull destacar el **Pere-Joan Cardona** per la confiança que va dipositar en mi obrint-me la porta a treballar al **CMCiB** (Centre de Medecina Comparativa i Bioimatge de Catalunya), una experiència impagable. Al centre m'han

acompanyat persones magnífiques que han fet la feina més distreta, la Yaiza, la Marta, la Sara, el Jordi, el Jorge, la Rosa, l'Osvaldo, el David... Gràcies per les col·laboracions i la companyia.

També esmentar l'enorme plaer de col·laborar amb tots els estudiants que m'han acompanyat, en especial l'Aida i la Imma, amigues i companyes. I als altres estudiants que ens han ajudat en aquest període de pandèmia, entre ells el Tomás, el Dani, el Pablo i la Cayetana; i aquests últims dies la Núria i la Marta que ho han fet tot plegat més divertit.

Aquesta tesi i aquests temps m'han permès iniciar col·laboracions fantàstiques que aspiro seguir en el futur. Em sento un privilegiat i estic molt content d'haver conegut persones tan competents: el Dani, l'Albert, la Teresa, l'Octavi, el Manolo, la Cristina, el Borja, el Didier, l'Ermengol, el Toni i molts altres grans científics d'aquest país.

Per acabar, dir que a nivell personal agraeixo la paciència que han tingut amb mi els meus pares, Maria i Jordi, i la germana, la Roser, a qui agraeixo els consells en disseny, i com no, a tots els amics i amigues, en especial el Ramon, el Xavi i el Francis amb qui he conviscut i l'Eli per ajudar-me sempre.

Aquesta tesi no hauria estat possible sense cap de les persones i equips esmentats anteriorment, i ben segur que m'he deixat algú. Gràcies a tots.

Preface

This thesis is about tuberculosis (TB) and coronavirus disease 2019 (COVID-19), two infectious diseases with a great impact on our society. As of late 2020, convincing someone about the importance of COVID-19 is not difficult because, after several lockdowns and restrictions with all their consequences, the European population is fully aware. But, can we say the same thing about tuberculosis?

During 2019, according to World Health Organization (WHO) estimations, approximately 55.4 million people died. Most of deaths were caused by non-communicable diseases (NCD), some 74% of the total amount. These diseases cannot be transmitted between individuals. The main NCD causes of death are cardiovascular diseases with 18 million deaths every year and cancers with nearly 10 million deaths each year. Eighteen percent of worldwide deaths (approximately 10 million cases) were caused by communicable diseases (CD). These are diseases that can be transmitted between individuals, most of them caused by viruses and bacteria. Some of the most important pandemic CD diseases are: flu, HIV/AIDS, hepatitis A, hepatitis B, zika and tuberculosis. In figure 1A the distribution of deaths between communicable and non-communicable disease is shown. These deaths are potentially avoidable by interrupting the transmission of these disease. Tuberculosis is the communicable disease that has caused the most deaths during human history. In 2019 TB was the main cause of death due to a single infectious agent.

The COVID-19 pandemic that started in late 2019/early 2020 distorted the counting of deaths during this year because, initially, COVID-19 was circulating but most deaths were assigned to other causes. The reported COVID-19 deaths are lower and the real numbers will probably never be determined. As may be seen in Figure 1B the number of estimated TB deaths in 2019 and the reported COVID-19 deaths are comparable. Both diseases caused between 5 and 10% of worldwide deaths in the years in question.

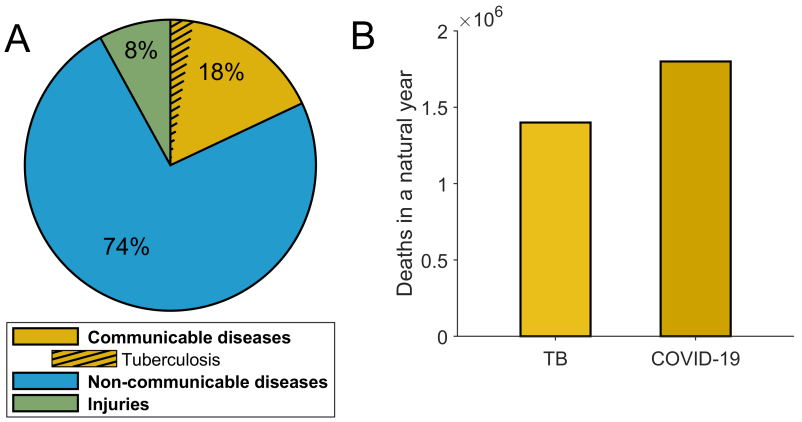

Figura 1: Causes of death. (A) Worldwide 2019 causes of death. (B) Comparison between estimated tuberculosis deaths in 2019 and reported coronavirus disease 2019 deaths in 2020. Data obtained from [182, 125].

Starting point

This thesis started 2 years before its formal initial point. During 2014-2015, in the second semester I did my final degree project of Physics Engineering "Bubble model, a computational model designed to describe active tuberculosis disease in mice". In this work, the Bubble model, a model that described active TB (ATB) lesions in mice [140], was adjusted to latent tuberculosis infection (LTBI) lesions in minipigs [36]. This project was directed by Dr. Clara Prats (one of my thesis supervisors) and Dr. Cristina Vilaplana with the collaboration of Dr. Pere-Joan Cardona and Dr. Daniel López. That experience was crucial to motivating me to write the present thesis.

My bachelor's thesis was followed by two master's theses. One, in 2016, "A 3D computational model for understanding tuberculosis lesions dynamics in lungs" (Advanced Physics Master) was supervised by Dr. Clara Prats and Dr. Ignacio Pagonabarraga. The aim of the thesis was to continue developing the Bubble model to explore the transition to ATB [34]. The second thesis was carried out in 2017 on "Machine learning to optimize an epidemiological tuberculosis model" (Master's in Bioinformatics and Biostatistics) supervised by Pau Andrio and with the collaboration of Dr. Clara Prats and Dr. Daniel López. The aim of that thesis was to adjust an heavy-epidemiological tuberculosis model to a set of parameters using machine learning [35].

Part of the work carried out during these years was the crucial background to starting the present thesis. The thesis formally started in October 2017 under the supervision of Dr. Sergio Alonso and Dr. Daniel López. Some months later Dr. Daniel López was replaced by Dr. Clara Prats. Pere-Joan Cardona and Cristina Vilaplana, medical doctors at the Experimental Tuberculosis Unit in Germans Trias i Pujol research Institute (UTE-IGTP), also played a fundamental role in supporting the medical and experimental part. After the appearance of COVID-19 Dr. Enric Alvarez-Lacalle also came to play a role.

Structure of the thesis

This thesis is a compendium of articles and is organized into 4 parts: Introduction, Articles, Conclusions and Appendix.

The Introduction consists of several chapters. Firstly, a general chapter that contains the basic framework is introduced. This part explains the most relevant issues about TB and COVID-19 (history and current situation, natural history and symptoms, treatments and vaccines and, finally, the open questions and challenges). Then, some of the different models that exist are introduced, focusing on the ones that are explained in this thesis and on their COVID-19 and tuberculosis applications. Finally, we state the goals and main objectives of this thesis.

The second part is the core of the thesis, as it contains the 5 articles. For each article, there are four sections: (1) a brief summary explaining the main points and findings of the article; (2) the main contributions in a bullet list; (3) applications and/or further work of the article explaining the applications in COVID-19 and further modifications that must be made or that have already been made; and, finally, (4) the article and its online link is included. As for the supplementary information, only links are included.

The third part of the thesis contains the main findings and the specific conclusions of the thesis, the perspectives for further work, and the bibliography.

Finally, there is there are two appendixes. Appendix A contains the glossary and appendix B a list of publications and research activity performed during the preparation of the thesis.

Índex

Part I.

Introduction

1 Basic framework

1.1 Epidemics and pandemics

An infectious epidemic is the rapid spread of an infectious disease to a large number of people in a given population within a short period of time. There is not a common value to determine if a certain number of cases is an epidemic or not, it depends on each disease baseline number of cases.

An epidemic of an infectious disease can be caused by several factors, among which are a change in the ecology of the host population, a genetic change in the pathogen reservoir, and the introduction of an emerging pathogen to a host population. Generally, an epidemic occurs when host immunity to either an established pathogen or newly emerging novel pathogen is suddenly reduced below that found in the endemic equilibrium and the transmission threshold is exceeded [26].

An epidemic may be restricted to one location; however, if it spreads to other countries or continents and affects a substantial number of people, it may be termed a pandemic [60]. The declaration of an epidemic usually requires a good understanding of a baseline rate of incidence; epidemics for certain diseases, such as influenza, are defined as reaching some defined increase in incidence above this baseline [75]. A few cases of a very rare disease may be classified as an epidemic, while many cases of a common disease (such as the common cold) would not. An epidemic can cause enormous damage through financial and economic losses in addition to impaired health and loss of life.

The World Health Organization (WHO) defines different phases of a pandemic [127]. The first three phases are before human-human transmission. Then, phase 4 is when a human-human transmission is verified, but cases are located in a single country. When they affect more than one country we say that they are in phase 5, and when the cross the boundaries of a WHO region it is said to be in phase 6.

Throughout human history, there have been countless pandemics. The most deadly pandemic in recorded history was the Black Death (The Plague), which killed an estimated 75–200 million people in the 14th century [187]. The last worldwide pandemics was the 1918 influenza pandemic (Spanish flu) with dramatic consequences [23]. Currently there are two pandemics declared, COVID-19 and HIV/AIDS [149, 184].

1.2 TB and COVID-19 comparison

Tuberculosis (TB) and coronavirus disease (COVID-19) are two infectious diseases that affect (mostly) the lungs and have some common symptoms like cough, fever and fatigue. Both diseases are transmitted by air though droplets. Common measures are applied to avoid their transmission such as the use of a mask and the contact tracing of infected individuals. Both diseases pose significant burdens; it is estimated that 1800 million people are infected by TB [125], and although at the beginning of 2020 there were just 41 reported COVID-19 cases, in less than a year and a half 160 million cases have been reported [64, 182]. The number of deaths reported in 2020 due to the two diseases is similar. In this sense, they put great stress on health systems and have devastating consequences. Also their economic consequences are huge; millions of dollars are spent every year on doctors, medicines, quarantines, and other derived costs. For example, it is estimated that the economic burden between 2006 and 2015 due to TB for twenty-two high-burden countries was $3.4 trillion [97]. Also COVID-19 pandemic consequences, just in the United States, are estimated to be around $16 trillion [53].

Despite these similarities there are enormous differences [174]. One of the main differences is that TB is caused by a bacterium, *Mycobacterium tuberculosis* (*Mtb*), and COVID-19 is caused by a virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Tuberculosis is a bacterial disease and COVID-19 is a viral disease with all the differences that this entails, especially in treatment [118]. It is important not to supply antibiotics to a viral infection to avoid proliferation of antibiotic-resistant germs, and moreover no potential benefit is obtained.

Coronaviruses diseases were observed in 1920-30 in chickens [153]; human coronaviruses were not observed until the 1960s [167], and the first outbreak was detected in China in November 2002 [7, 128]. The first detected case of COVID-19 was reported on the 31st December 2019 [126]. In contrast, tuberculosis has cohabited with humans for thousands of years. It is estimated that this coexistence started 70,000 years ago [21, 50, 20] an enormous difference compared with coronaviruses. Despite TB's historical importance, nowadays, COVID-19 dominates scientific literature and media, and TB is normally considered an ancient disease, but it should not be neglected.

Related with this time scales differences, TB has been widely studied whereas COVID-19 investigation just started in 2020. Then there is a

system to estimate TB cases and deaths worldwide, which is why when we speak about TB cases and deaths there is an estimated quantity from the WHO [125, 183]. In COVID-19 this system is not yet implemented and when speaking about COVID-19 cases and deaths they are always the numbers reported. In the future a system to estimate COVID-19 cases and deaths may be implemented [64, 182].

Progression of the two diseases is very different. On the one hand COVID-19 has a fast progression. Its incubation period is of around 5 days [112] and in less than a month after initial infection [186] most patients are either recovered or dead. There are also some complications that can last for months [106, 105, 84]. COVID-19 long-term complications are not fully evaluated, since it has been less than two years since the first reported case. In contrast, TB has a slow progression. Active tuberculosis (ATB) can be developed years after initial infection. In half of the cases it is during the two years after initial infection, but it can appear till 7 years after initial infection [88]. In fact, on average there is a time lapse of 60 days from when a person starts to develop symptoms until they are diagnosed, even in high-income countries [156]. Individuals suffering active tuberculosis disease can last for a long time; deaths are reported after one or two years from diagnosis [32]. Then, the time scales of the two diseases are very different, days for COVID-19 and years for TB.

As may be seen in Figures 2 and 3, the incidence of cases and deaths from TB and COVID-19 are very different. In Figure 4 there may be seen the distribution of cases and deaths per continent. Tuberculosis mostly affects 2 continents (Africa and Asia) that contain 96% of cases and 97% of deaths, respectively. In fact, the countries most affected by TB are those with lower human development index (HDI) and gross domestic product worldwide [142, 49]. COVID-19 has a more expansive distribution and affects all continents similarly; the continent with the most cases and deaths reported in 2020 was Europe [183], which is probably why greater attention is paid to COVID-19 compared to TB.

In fact, there is a positive correlation between HDI and the number of COVID-19 reported cases [42]. This correlation may be explained by several factors: different diagnostic capacities, different age (then susceptibility), racial differences [169], different social relations, and other factors. Figure 5 shows the correlation between reported cumulative COVID-19 incidence up to 1st January 2021 [64, 182], and the last twenty years

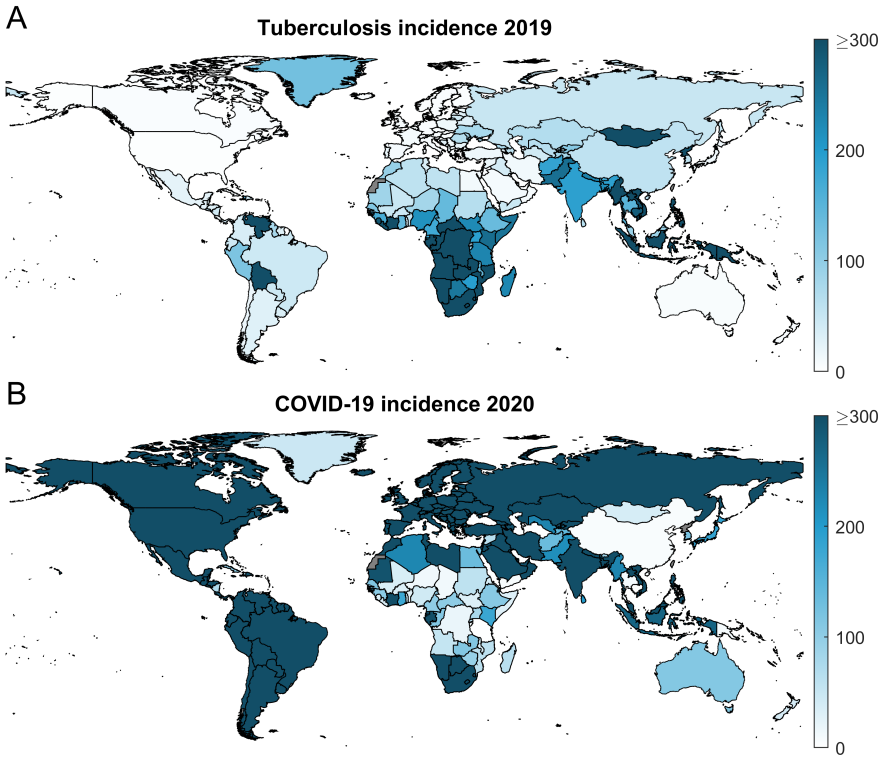

Figura 2: Tuberculosis (TB) and coronavirus disease 2019 (COVID-19) world case incidence. (A) Each country color depends on TB 2019 estimated case incidence as blue legend bar. (B) Each country color depends on COVID-19 2020 reported case incidence as blue legend bar. Countries with no available data are grey. COVID-19 data were obtained from [64] and TB data from [183].

estimated cumulative incidence of TB [183], gross domestic product per capita [49], mean age [48] and human development index [142] for 227 world countries and territories. There is seen to be a positive correlation between GDP, HDI and mean age. This is an expected result considering that quality of life and lifespan are strongly related with GDP. We may observe that COVID-19 cases correlate positively with GDP, Age and HDI, while TB correlates negatively. Figures 2, 3, 4, and 5 probably explain why although the two diseases have a very similar death impact in the world population, popular attention, the media, and scientific reports cover them so differently. By simplifying, COVID-19 is a disease of the rich and TB is a disease of the poor.

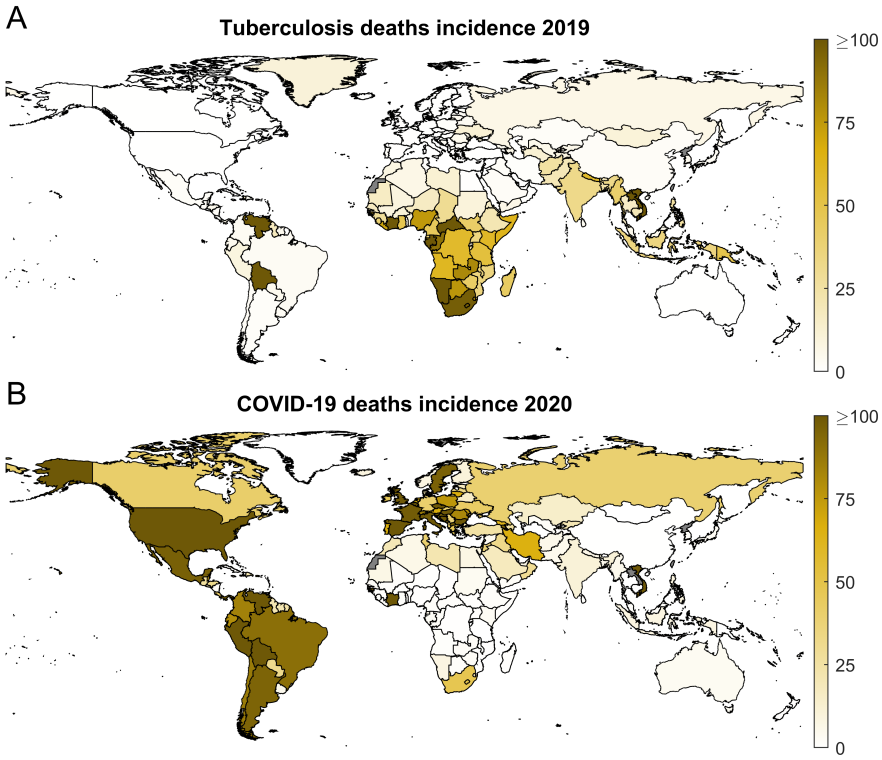

Figura 3: Tuberculosis (TB) and coronavirus disease 2019 (COVID-19) world deaths incidence. (A) Each country color depends on TB 2019 estimated deaths incidence as yellow legend bar. (B) Each country color depends on COVID-19 2020 reported deaths incidence as yellow legend bar. Countries with no available data are grey. COVID-19 data were obtained from [64] and TB data from [183].

Some authors also suggest that latent tuberculosis infection (LTBI) can give the host some protection against COVID-19 and this could explain the negative correlation between the two diseases [69]. Further analysis will determine the reason for these negative correlations. The main hypotheses are: (1) LTBI protects against COVID-19, (2) low diagnostic level distorted COVID-19 detection in countries with low income, (3) lifestyle in less developed countries helped to protect people from the COVID-19 pandemic, and (4) age determines susceptibility to severe infection and less aged populations are protected.

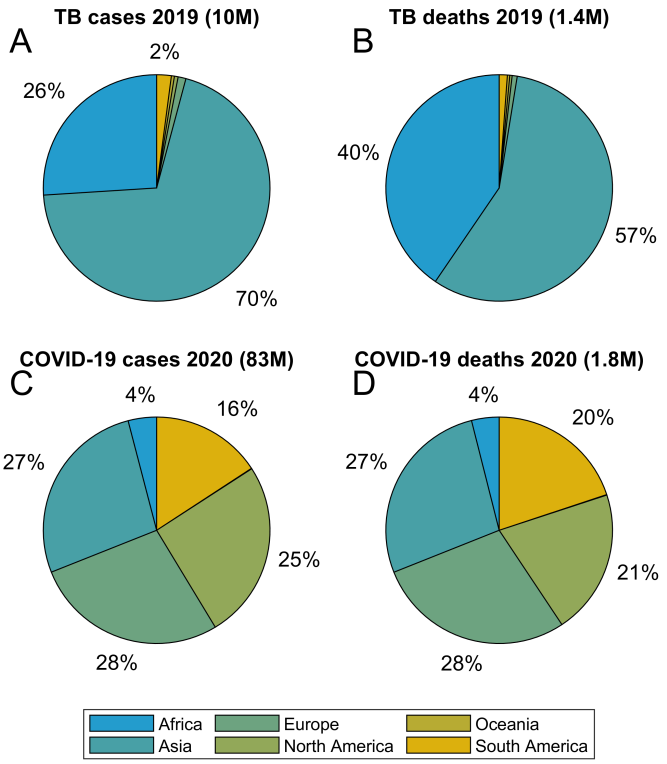

Figura 4: Tuberculosis (TB) and coronavirus disease 2019 (COVID-19) cases and deaths distribution by continents. (A) Distribution of TB estimated 2019 cases per continents. (B) Distribution of TB estimated deaths per continents. (C) Distribution of COVID-19 reported 2020 cases per continents. (D) Distribution of COVID-19 reported 2020 deaths per continents. Continents colour legend can be seen at the right. COVID-19 data were obtained from [64] and TB data from [183].

To attach numbers to the differences explained in 2019, the number of estimated deaths due to TB was 1.4 millions [125] and 1.8 million was reported as due to COVID-19 [64, 182]. Eighty-seven percent of TB cases were detected in the 25% of countries with the lowest gross domestic product, representing 58% of worldwide population and 37% of total COVID-19 cases. In contrast, 25% of worldwide gross domestic product is owned by just 7% of the population; in those countries 30% of COVID-19 cases were detected, but less than 0.5% of TB cases. We can observe that there are great inequalities in TB cases, COVID-19 cases, and population distribution. Scientific efforts are not comparable, with

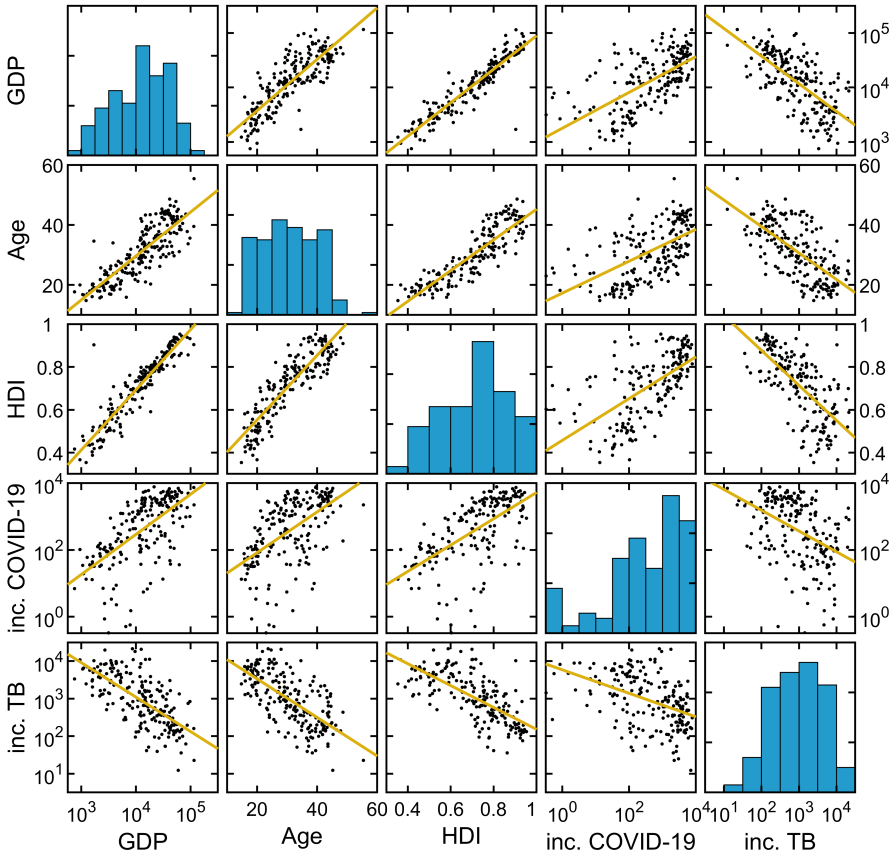

Figura 5: Correlations between gross product per capita, mean age, human development index, coronavirus disease 2019 (COVID-19) cumulative incidence in 2020, and tuberculosis (TB). Data from 227 countries are used. Diagonal plots are histograms of different values: Gross domestic product per capita (GDP), mean age of the country population (Age), human development index (HDI), cumulative reported COVID-19 incidence from 1 January 2020 to 15 May 2021 (inc. COVID-19) and cumulative estimated TB incidence from 2000 to 2019 (inc. TB). In non diagonal plot each point represents the data from a country, yellow solid line shows a linear fitting. GDP, inc. COVID.19 and inc. TB are displayed in logarithmic plot. Data obtained from [49, 48, 142, 182, 183].

a simple Pubmed search we find that since 2020 there are 90874 entries about COVID-19 [120] and only 9904 about TB [121]: similar numbers of deaths and a tenfold difference in scientific coverage.

1.3 Physiological description of human lungs

Tuberculosis and coronavirus disease 2019 mainly affect the lungs. In TB, in 80% of the cases the lung is the only affected organ; all models that we are going to consider are from pulmonary TB. In COVID-19, the lungs are not the only affected organ; COVID-19 also affects kidney and liver. Most of the visible COVID-19 consequences are observed in the lungs. The studied COVID-19 models presented in this thesis are on the population scale and not on the organ scale.

The lungs are the essential respiratory organ in humans. Their function is to supply oxygen to the bloodstream and in a gas exchange release carbon dioxide to the atmosphere. A large surface is used to perform this exchange between blood and air. The smallest lung unit where this exchange takes place is the alveolus, a thin walled air sac formed by specialized cells.

Human lungs occupy a global volume of around 5 liters (female lungs are smaller). Lungs contain around 700 million alveoli, each alveolus having an small radius of around 0.15 mm. Lungs are surrounded by a 0.3 mm pulmonary membrane, the pleura [178]. Lungs are divided into lobes; in humans two in the left lung and three in the right lung, but these numbers may vary between species. Lobes are surrounded by pulmonary membranes. They are also divided into segments; the smallest segment division is the secondary lobe. Secondary lobes are also surrounded by a thinner pulmonary membrane.

Secondary lobes are around 1 cm^3. They contain acini which are groups of alveolar sacs made up of clusters of alveoli. Figure 6 is a schematic representation of human lungs divisions from the trachea to the alveoli.

Air enters through the mouth to the trachea that supplies air to the millions of alveoli. Air distribution is produced through the bronchial tree that is formed by the trachea and its successive divisions. The trachea divides into two main bronchi that enter both lungs. The bronchi continues to divide within the lungs, and after multiple divisions (around 30) reach the terminal bronchioles which are connected to alveolar sacs [178].

The lung environment is very moist which makes it hospitable for bacteria and viruses. This fact and taking into account that it is an

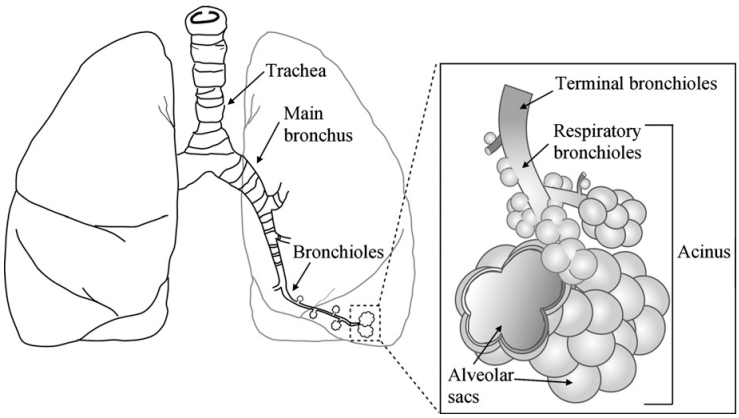

Figura 6: Human lungs. Schematic diagram of the bronchial tree divisions and their ramification into alveolar sacs. Obtained from [162].

entry point to the body mean that the lungs are the target of different pulmonary infectious diseases such as tuberculosis, coronavirus disease 2019, and pneumonia.

2 Tuberculosis

2.1 History and current situation

Tuberculosis is a major threat to humankind. Indeed, it has been estimated that this disease has caused 1,000,000,000 deaths in the last 200 years [132]. Since the world emergency declared by World Health Organization in 1992 [129] the efforts made to control TB expansion have increased in recent years. Despite these efforts, intended to improve prevention, diagnosis and treatment, TB is still a challenge and its incidence is only declining slowly, causing 1.4 million deaths in 2019 and 39 million in the last 20 years [125].

In Figure 7 we may observe the TB cases and deaths in the recent years. As may be seen there exists a declining tendency but we are far from our goals. This declining tendency may now be truncated due to the COVID-19 pandemic; in fact, the WHO estimates that there has be a 12-year retreat in the TB eradication plan [180, 125, 164]. Partial reporting data of 84 countries estimates a 21% decrease in diagnosed cases. In recent years the percentage of detection was increasing and this reduction represents a significant setback. Another side effect is that 1.4 million people did not receive treatment due to the COVID-19 pandemic. There is a well-founded fear that the consequences of the COVID-19 pandemic in the TB epidemic could be dramatic [180].

The WHO estimates that nearly 10 million people developed TB in 2019. In fact, they also estimated that between one third and one quarter ($\sim$1.8 billion people) of the worldwide population are already infected with *Mycobatectium Tuberculosis*, the bacteria that is responsible for TB [125]. Between 5 and 15% of infected people will develop an active TB disease in a few years' time [91]. It is estimated that in the coming years between 90 and 270 million people will develop ATB.

Normally, TB evolution is slow; it can last from a few months to some years. Nowadays only 1.4% of humans that suffer from this disease die from it. In treated individuals this is reduced to nearly zero [25]. But if TB is untreated it is estimated about 3% or more of them will die [160]. Moreover, some populations with low immunological response or people that suffer from other lung diseases are more susceptible to an acute and serious illness. This includes children, the elderly, HIV-positive people, cancer patients, diabetics, smokers, and alcoholics, among others [11].

Several groups have attempted to determine the origin of *Mycobacterium tuberculosis* in order to better understand its highly virulent nature. Originally it was asserted that its origin is a zoonotic infection evolving from an ancestor of *Mycobacterium bovis* [158], but current data support the opposite. Thus, *Mtb*'s most recent common ancestor emerged around 73,000 years before the common era, originating specifically in the so-called anatomically modern humans (*Homo sapiens*) from an environmental mycobacterium [21, 50, 20]. Then, *Mtb* has cohabited with humans for more than 70 thousand years.

On March 24, 1882, Pr. Robert Koch discovered the bacteria which was causing TB. *Mtb* can also be referred to as Koch bacillus due to its discoverer, and that's why March 24 is World Tuberculosis Day [102].

2.2 Natural history and symptoms

Tuberculosis is spread by air though airborne particles. This is the main reason why it mostly affects the lungs, but extrapulmonary TB has also been an uncommon pathogenicity that occurs in 20% of cases [131].

Tuberculosis infection starts when *Mtb* arrives at a pulmonary alveolus and it is phagocyted by an alveolar macrophage. Generally, bactericidal macrophage mechanisms are not enough to eliminate them. In fact, *Mtb* is able to multiply inside the phagosome, thus avoiding macrophage apoptosis [16]. Finally, when there are between 32 and 64 bacilli, they cause macrophage necrosis and thereby enter the extracellular milieu, where they are phagocyted by other macrophages which also fail to control the bacillary growth and are likewise destroyed. This macrophage elimination triggers the inflammatory response and macrophages, neutrophils, and other cells are directed to the infection point. Inflammatory response is not able to contain bacilli growth and, in fact, it accelerates the exponential growth of bacilli.

This cycle of exponential bacillary growth ends once the specific immune response appears. Immune response is able to contain this progression and finally leaves a calcified lesion formed by different types of cells. Growth containment is also helped by an encapsulation process led by fibroblasts from the pulmonary membranes (mainly secondary lobe membranes), surrounding the lesions and preventing their growth. According to the Dynamic Hypothesis from Cardona [28] there is a certain possibility that few bacilli will escape from the lesion, mostly inside

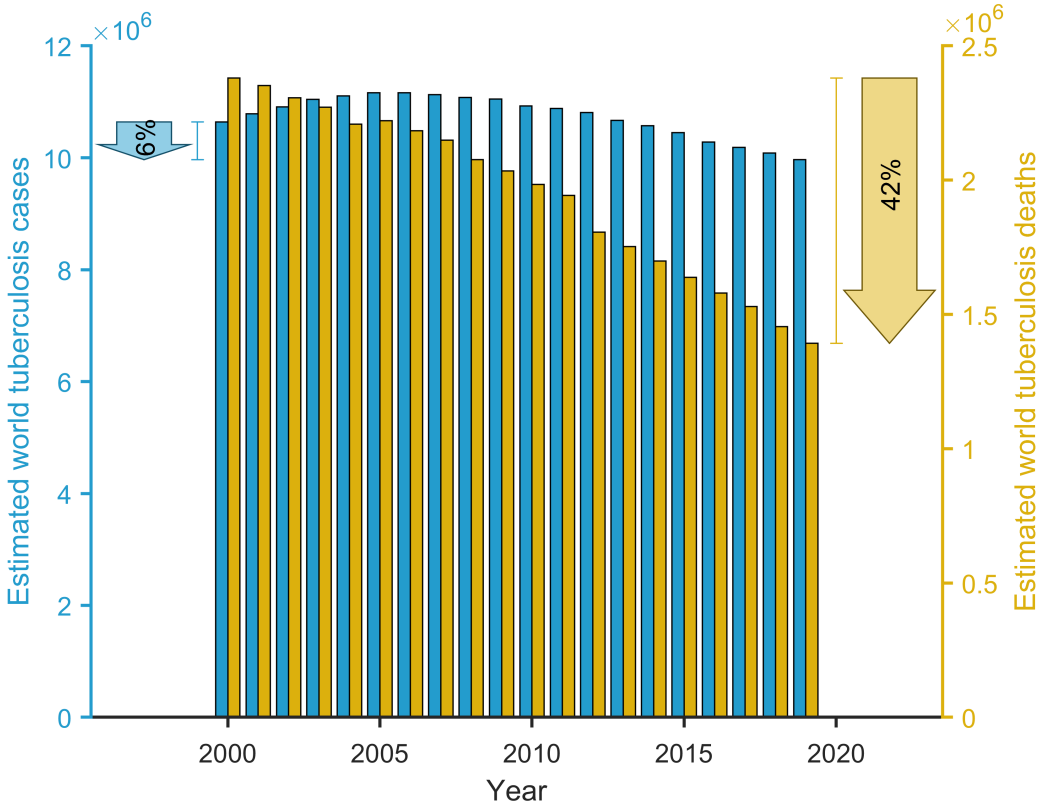

Figura 7: World tuberculosis (TB) cases and deaths, evolution from 2000 to 2019. In blue the estimated TB cases worldwide each year, with a 6% decrease in this 20-year period. In yellow the estimated TB deaths worldwide each year, with a 42% decrease in this 20-year period. Data obtained from the WHO [183].

a foamy macrophage, and start a new infection in other alveoli. This is known as the endogenous reinfection process, and it takes place through the bronchial tree. Other authors also suggest that this reinfection is not produced [4] or is produced though other paths in homogeneous ways [9].

If the host is immunocompetent, in most cases the infection is controlled and a calcified lesion remains. This host is said to be infected but not ill because it does not show symptoms and cannot transmit the disease. It can remain in that state without being aware of the infection and not developing visible consequences. According to the Dynamic Hypothesis, there is a certain possibility that the infected foamy macrophage will

go to the upper lobe. In this location the conditions favor extracellular growth of bacilli (liquefaction occurs in the intragranulomatous necrosis) and lesions easily become uncontrolled. In fact, most active tuberculosis lesions are observed in the upper lobes [17, 170].

Success in the control of the lesion depends on a correct equilibrium between the inflammation response that promotes the growth of the lesion (due to accumulation of different types of cells) and the immune response that stops and controls the growth. If one of these two responses has an incorrect behavior the infection can evolve to disease.

To sum up, there are three types of people with respect to their TB status: healthy people, who do not have bacilli nor are sick; infected people, who have bacilli but their lesions are controlled, they do not present symptoms and they cannot infect others, they have what is called a latent tuberculosis infection; and sick people, who have bacillus inside and the capacity to infect other people, and have an active tuberculosis disease, presenting symptoms. A fraction of them can die if not properly treated.

It has been observed that some TB complications are associated with the appearance of sputum in the lungs. In fact, sputum smear test is used to determine its presence and diagnose. Patients diagnosed with positive sputum smear have a greater probability of developing a more severe disease; in fact, it is used as a severity indicator to detect patients with possible complications. This severe sign increases mortality 15 times with respect to those who do not present this complication [146]. Also it has been seen that patients with smear negative TB are less infectious than patients with positive sputum smear TB [130].

2.3 Treatments and vaccines

Each person with pulmonary TB disease can infect between 10 and 15 people per year, on average [161]. However, it can be prevented and cured if taken seriously as a threat to individual and public health. Prevention, diagnosis, and effective and appropriate treatment are required to prevent the disease spreading.

Effective drug treatments were first developed in the 1940s. The most effective first-line anti-TB drug, rifampicin, became available in the 1960s. The currently recommended treatment for new cases of drug-susceptible TB is a six-month regimen of four first-line drugs: isoniazid,

rifampicin, ethambutol, and pyrazinamide. Treatment success rates of 85% or more for new cases are regularly reported to the WHO by member states. Treatment for multidrug-resistant TB (MDR-TB), defined as resistance to isoniazid and rifampicin (the two most powerful anti-TB drugs) is longer, and requires more expensive and more toxic drugs. For most patients with MDR-TB, the current regimens recommended by the WHO last 20 months, and treatment success rates are much lower.

For the first time in four decades, new TB drugs are starting to emerge from the pipeline, and combination regimens that include new compounds are being tested in clinical trials. In fact, recent results suggest that the treatment period could be reduced to four months [61]. This is great news because as the treatment is expensive and long. It is often abandoned in some countries, mostly in poor and lower income countries.

There is also one TB vaccine that has been in use since 1921. The bacillus Calmette-Guérin (BCG) vaccine is used as part of national vaccination programs in countries with many cases of TB. BCG is used in children and special cases and does not protect from pulmonary disease caused by *Mtb*, nor does it prevent latent TB infection from progressing to active disease. It does, however, prevent some serious TB complications in children, such as TB meningitis. The vaccine is generally not used in adults, and the vaccine in children does not prevent spread of the disease. Many researchers are working to develop a more effective tuberculosis vaccine. The hope is to develop a vaccine that prevents infection with tuberculosis, which would reduce the great global burden of disease and also reduce transmission of the *Mtb*. There are several TB vaccines in Phase I or Phase II trials. For the time being, however, a vaccine that is effective in preventing TB in adults remains elusive [125].

Tuberculosis is present everywhere, although rooted in developing countries. The incidence of TB caused by *Mycobacterium tuberculosis* strains resistant to anti-TB drugs (MDR-TB) has increased due to poor administration of antibiotics. In fact, some strains that are untreatable have been identified. Nowadays the WHO estimates that 33% of new TB cases are due to MDR-TB. The highest proportion of MDR-TB cases is found in countries of the former Soviet Union [125].

2.4 Challenges and open questions

In the coming years there are a lot of challenges that we must face to eradicate and minimize the impact of TB in the world. The most important one in the short term is to help the health system recover from the COVID-19 pandemic and to resume the path to the End TB strategy. This strategy will need political and financial commitment to reach TB control. The WHO estimates that there has been a setback of 12 years in TB control due to the COVID-19 pandemic [180, 164, 125].

Important challenges are to improve treatments as they are expensive and long and to face the problem of multidrug-resistant strains that increase every year. In order to design better treatments or find better vaccines, we need to better understand natural history and lesion formation. The key point in this sense is to understand the processes, mechanisms, and characteristics that make some patients evolve from a latent TB infection to an active TB disease.

In this thesis we are going to focus on the study of three challenges. First, we are going to study how lesions are formed by building a model that reproduces TB's known natural history at the alveolus level, in order to determine the mechanisms that are important in infection control. Encapsulation process has an important role in lesion control and implementing it is useful to improve our understanding of how this process works.

Secondly, we are going to explore the dynamic tuberculosis hypothesis in a computational minipig model as a possible evolution of TB lesion infection. We are going to build a model that reproduces experimental LTBI data, and later we will use this model to explore the parameters that cause this LTBI to evolve to an ATB, trying to relate the parameters with biological mechanisms, since the factors and mechanisms that drive an LTBI to an ATB are not fully understood.

Finally, we are going to study the coexistence of TB with humans for 70,000 years and the factors that conditioned life in the past. In fact, we are going to see that TB conditioned life a long time ago. One key factor that has not been explored yet is the role of female resistance and tolerance so as to understand the consequences on TB incidence and evolution.

3 Coronavirus disease 2019

3.1 History and current situation

COVID-19 is a viral disease caused by severe acute respiratory syndrome coronavirus 2. Coronaviruses take their name from the distinctive spikes with rounded tips that decorate their surface, which reminded virologists of the appearance of the sun's atmosphere, known as its corona.

We start with a short review of the COVID-19 pandemic' timeline [181]. On 17 November 2019 a person presented the first detectable case of COVID-19 in China. But it was not till 31 December 2019 that the WHO' country office in China reported a cluster of 41 patients with a "viral pneumonia" in Wuhan. On 9 January 2020 Chinese authorities determined that the outbreak was being caused by a novel coronavirus, and two days later the first death due to this new disease was reported. On 13 January 2020 the first case outside China was detected in Thailand; the virus was circulating between countries without countries being able to prevent it.

Expansion continued and on 11 March a world pandemic was declared by the WHO with 121,172 cases, most of them (80,115) in China but present in 117 countries around the world. On that day this new virus had caused 4,368 deaths to date in 26 different countries, most of them in China [64, 182]. It was too late to control the expansion with soft measures and countries were forced to close firstly their borders and later keep their citizens at home. Non-pharmaceutical interventions (NPI) were the only solution that countries had in the short term because there was neither a vaccine nor a treatment that could contain the effects of the virus.

On 4 April there were reported to be more than one million cases worldwide. Cases multiplied by a factor of 10 in less than a month. This tendency continued in following days, and, by July 2020, there were reported to be more than ten million cases and half million deaths. The consequences of the pandemic were devastating. Figure 8 shows the evolution in the number of cases (A) and deaths (B) from 22 January 2020 until 15 May 2021.

While cases and deaths continued their unstoppable rise, the first hope was reported on 2 December 2020, after 64 million cases and 1.5 million deaths, when the United Kingdom authorized a Pfizer-Biontech

vaccine. It was the first country in the world to have a clinically approved vaccine in supply. Nowadays several vaccines have been approved by regulators worldwide, and more are under investigation. In the past half year, vaccination velocity has grown exponentially as can be seen in Figure 8C.

During the pandemic, science and knowledge have evolved very rapidly. In fact, scientists the world over have flooded journals with an overwhelming number of articles. Also there has been a an explosion in the number of preprints, which are non-peer-reviewed publications. Non peer-reviewed publications facilitated the rapid communication of new findings, models, treatments... But also some errors were committed, the most famous being using hydroxychloroquine as a treatment [114, 101]. Nowadays the main parameters of COVID-19 (incubation period, time from onset of symptoms to death, susceptibility, death probability...) are known or there is a consensus about them [112], whereas at the pandemic outset very little was known.

On 12 May 2021 there were 760,000 new cases reaching over 160 million cumulative cases, 14,000 new deaths with a total reported 3.3 million deaths, and 20 million of vaccine doses administrated, reaching a total of 1.37 billion doses administered. 500 days after the first case was reported, we have several approved vaccines that in 5 months reached 8.7% of the world population. Vaccination changed the COVID-19 dynamics and lot of countries are moving towards a normal life; in fact, many countries are removing or relaxing their control measures. Unfortunately, most of the vaccines are not equally distributed with most of the vaccines administered in rich and more developed countries [110, 123]. In the future, if the hot spots are in poor and less developed countries, this situation is going to continue many years. There is the risk that a similar situation to that of TB will be created. This is a long-term hypothesis, and nowadays COVID-19 will continue to be a big challenge for the worldwide situation.

In Catalonia as of 12 May 2021 there were reported to be 14,000 deaths due to COVID-19 [31], plus 8,000 suspected deaths, 615,000 cases, and 3.3 million administered vaccine doses [45].

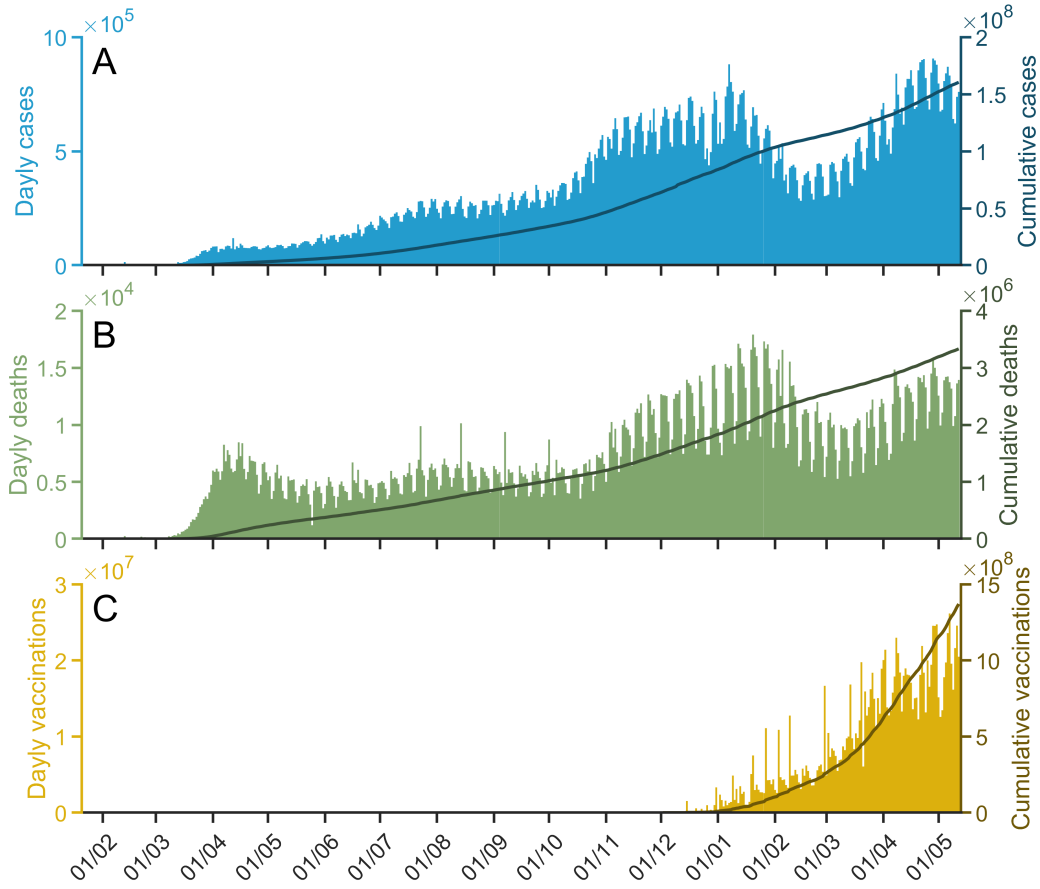

Figura 8: World evolution of coronavirus disease 2019 (COVID-19) cases, deaths and vaccinations from 22 January 2020 to 15 May 2021. (A) Blue bars for the daily number of reported COVID-19 cases worldwide, on the left axis. Darker line for the cumulative count, on the right axis. (B) Green bars for the daily number of reported COVID-19 deaths worldwide, on the left axis. Darker line for the cumulative count, on the right axis. (C) Yellow bars for the daily number of reported COVID-19 vaccinations administered worldwide, on the left axis. Darker line for the cumulative count, on the right axis. Data obtained from [64, 182, 110].

3.2 Natural history and symptoms

The incubation period for COVID-19 is up to 21 days, but the median time is about 4-5 days from exposure to symptoms onset [112, 96]. One of COVID-19 main challenges is that contagiousness starts before symptom onset [52]. Some 97.5% of people with COVID-19's who have symptoms will do so within 12 days from SARS-CoV-2 infection. That is a fast progression compared for example to TB [112]. The time between onset of symptoms and deaths is estimated at around 10-20 days for COVID-19 [186, 56, 100].

An interesting challenge is that there is an unknown but important percentage of people that are asymptomatic (depending on the study, ranging from 20% to 90%, with between 40% and 50% being the most widely accepted values) [134, 31, 95, 144]. Which is their role and the transmission and pandemic evolution is not a clear fact yet.

The signs and symptoms of COVID-19 presenting at illness onset vary, but over the course of the disease many people with COVID-19 will experience some of the following symptoms: fever, chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting and diarrhea [77, 56, 85].

As may be seen there is a wide range of possible symptoms; this was one of the difficulties in detecting the disease in the primary stages of the pandemic. There was not a clear symptom pattern to identify the possible cases. Symptoms may differ with severity of disease. For example, shortness of breath is more commonly reported among people who are hospitalized with COVID-19 than among people with milder disease (non-hospitalized patients)[159].

COVID-19's most common comorbidities [176, 78] have been identified. Chronic obstructive pulmonary disease and cerebrovascular disease with a hazard ratio of 6 and 4, respectively, are the most important ones. One of the most important factors in death rates is the age of the patients; globally it is estimated that the fatality rate is around 1%, but it ranges from around 30% in 90-year-old to a nearly zero probability in the 20-30 year-old range [10, 100, 138, 90]. Another important observed factor is gender; it has been estimated that there is an increase of between 1.5 and 2 in the probability of death between males and females [45, 143, 133, 68]. Hospitalization rates depend on the country (different health

systems, different hospitalization criteria...) but they range between 2 and 5% of cases. There is also a factor of hospitalization rates in men and in women [45, 143, 31].

3.3 Treatments and vaccines

Although some treatments exist [55, 57], they are not commonly used. In fact, in the United States the Food and Drug Administration (FDA) has only authorized one of them (Remdesivir [13]). In addition, there are also some treatments that can be administered in emergency cases (authorized for emergency use). In fact, vaccine development was much faster than treatment development. That is a common fact in viruses; it is usually more difficult to find a treatment than to develop a vaccine [66].

Vaccine development this last year has been very fast, and several vaccines have been approved and are being administered now. Usually a vaccine trial can last 10 years, but in this case, we were able to reduce it very much due to three main factors. First, techniques that have been used in other diseases could be adapted and safety protocols were evaluated faster; second, the clinical trials were not sequential but partially overlapped to reduce time, and preliminary results were presented to authorities so that the regulators could start reviewing them without waiting till the end; finally, the high incidence of COVID-19 sped-up Phase III trials. Three types of vaccines have been developed: (1) mRNA vaccines, which contain mRNA material from SARS-CoV-2 that gives cells instructions for how to make a harmless protein that is unique to the virus to be prepared if the individual is infected in the future; (2) protein subunit vaccines that include harmless pieces (proteins) of the virus that cause COVID-19 instead of the entire germ; (3) vector vaccines that contain a modified version of a different virus from the one that causes COVID-19. This genetic material gives cells instructions to make a protein that is unique to the virus that causes COVID-19 [46].

In Europe, as of May 2021, there are four vaccines authorised by the European Medicines Agency (EMA). These vaccines have shown high efficacy: 70.4% for Oxford-AstraZeneca vaccine [175], 95.0% for Pfizer-BioNTech vaccine [137], 66.9% for Johnson & Johnson [150], and 94.1% for Moderna [8] in clinical trials which tested schedules of two doses given 3 weeks apart (Pfizer), 4 weeks apart (Moderna), 4 or 12 weeks apart

(AstraZeneca), and in a single dose (Johnson & Johnson). Current studies are showing a good performance in "real" world studies; for example a 98% death reduction in nursing homes in Catalonia [25], reduction of transmission in Israel [71], and a reduction even after first dose administration [15, 171]. Nowadays,a total of 8 vaccines have been approved by different countries. In addition to the ones approved in Europe, the other four are: CanSino, Sinopharm, Sinovac and Sinopharm-Wuhan. In fact, there are more than 50 vaccines in phase I trials, and more than 30 have already started phase II and III as well [103, 57].

Initially, vaccine supply was slow and choosing a proper vaccination strategy was crucial. Most countries decided to prioritise by age, which is one of the most important factors that affect death rate. This was also due to logistics as it is easier to organize a vaccination by ages ranges [90, 100]. Also the people most exposed were vaccinated first, e.g. health workers [47].

In parallel, many experts have proposed modifying vaccination schedules, despite uncertainty surrounding the effects of using longer dosing intervals than those studied in clinical trials [115, 33].

3.4 Challenges and open questions

At the beginning of 2020 all questions about COVID-19 were open: what is COVID-19's transmissibility? Which is the incubation period of this disease? Are people equally susceptible? What are the events where most contagion is produced? Are schools a big focus and are kids superspreaders? How does immunity work? Can people be sick again? Nowadays, some of these questions are resolved as explained in the previous sections. But there are still a lot of open challenges, some of them explained below.

Asymptomatic COVID-19 is one of the current key points. It is known that there is a fraction of the population that is asymptomatic, but what is exactly this fraction? We can see some asymptomatic 90-year-old individuals and, at the same time, 30-year-old individuals with strong symptoms that end up in the ICU. Is there a way to determine the factors that condition the symptoms or it is just a matter of the viral load to which we are exposed? What does being asymptomatic imply? Does an asymptomatic infection entail less subsequent immunity or less transmissibility? A lot of articles are trying to solve these issues

but there is no clear consensus on them [185]. Nevertheless, scientific knowledge is advancing very fast and we will soon have answers to all these questions.

A very common topic these days in media is the new variants of COVID-19. The first variant that raised our attention was the so called British variant (alpha). Some reports and papers show that it is more transmissible [1, 109]. It was postulated that this would change the situation radically but in May 2021, although being the dominant lineage in Europe [58, 119, 19], the situation was mostly controlled in Europe thanks to vaccination and control measures. In Catalonia, the third wave was no worse than the second due to the emergence of this variant [45]. In fact, during the summer of 2020 there was also a substitution process of the Wuhan variant by a new one (not drawing as much attention as the British one) [82]. Since a virus is not controlled it is evolving and mutating, and new variants are appearing all the time. The most notable ones are South African (beta) [136], Brazilian (gamma) [124], and Indian (delta) [165]. Is it possible that some of these variants avoid immunity given by vaccines? This question is under study, and for, the moment, vaccines have shown that efficacy over different variants is maintained but with differing effectiveness [67, 27, 177]. While the virus is not controlled there is always the possibility that a new variant will totally change the paradigm.

Another recurring topic that has not achieved a scientific consensus is the role of children in the pandemic. Most countries closed schools [65] to avoid propagation, but recent studies suggest that it was unfounded. In fact, it has been seen that kids are equally or less susceptible and contagious than adults [74, 155]. In fact, they are a reflex of what is happening in the society [89, 37].

The most recurring challenge is the “crystal ball” needed to predict an outbreak before it starts. Different models have shown good capacity to predict the dynamics of one outbreak once it has started, but none of them has been capable of anticipating the onset of a new outbreak.

To understand the questions and challenges that we want to face in this thesis we must understand that in March 2020 we were blind and we had very little knowledge about this disease. We have centered this thesis on two challenges:

First of all, during the first months of the pandemic all over the world there was a testing limitation, and during the so-called first wave in Europe (March to May 2020) we could not detect all the cases [138]. In fact, most of the detected cases were diagnosed directly in hospital [31]. Then, how can we estimate the real incidence of COVID-19? This problem (to a lesser extent) is also present now because not all countries have the same detection rates [43]. We also want to determine at which point of the disease timeline we were detecting cases and what the detecting percentage was.

Finally, as explained, at the beginning of the pandemic we were blind in many aspects and we had many uncertainties about disease parameters (incubation period, transmisibility, relation between measures and transmisibility...) so it was not possible to build a mechanistic model. But can we predict the pandemic evolution without knowing all these parameters? With an empirical model, for example. This is the second challenge that we faced.

4 Models of an infection

A model is a simplified representation or description of a system. In science, models are necessary in order to (i) understand the studied system, (ii) make predictions about its probable evolution, and (iii) constrain or manipulate it to produce a desirable condition [81].

The use of models can be extended to different scientific branches. For example, in epidemiology, we need to (1) understand the dynamics of a certain disease, (2) make predictions about the probable evolution, and (3) condition the evolution of the epidemics under different circumstances, by means of control measures or vaccination campaigns. Moreover, in biomedicine, we aim to: (1) understand the dynamics of cells and systems involved in a certain disease, (2) make predictions about the probable evolution of a patient, given each patient's constraints, and (3) condition the evolution of the disease in order to control its symptoms and, if possible, cure it.

In [81] four kinds of modelling forms are stated:

- Conceptual or verbal models, which are descriptions in a natural language.
- Physical models, which are real, physical mock-ups of a real system or object.
- Diagrammatic models, which are graphical representations of the objects and relations.
- Formal or mathematical models, which are usually formalized using algebraic or differential equations.

In this thesis we are going to study TB and COVID-19 using mathematical models that are widely used in epidemiology and biomedicine. For example, using a mathematical model to predict an epidemic evolution, or to study how two diseases cohabit. One of the novelties in this thesis is that we are going to submit a conceptual model (hypothesis) for falsifiability using a complex mathematical model.

4.1 Animal models

In biology, we can basically talk about three kinds of experimental models, which can be defined as physical models according to the previous classification: *in vitro*, experiments made to occur outside the living organism in an artificial environment; *ex vivo*, experiments done in or on tissue from an organism but an external environment; *in vivo*, experiments occurring or carried out in a living organism (animals normally).

All of them are necessary steps in experimental research on TB, COVID-19, or any other disease. *In vitro* cultures are mainly devoted to understanding the biological features of *Mtb* and SARS-CoV-2. For instance, they are important to elucidate which environmental conditions are more favorable for its rapid growth.

The choice of the animal model to carry out such experiments is a key point of the experimental protocol design. The most frequently used animals as *in vivo* models are mice, since they are low cost compared to other species and relatively easy to handle. Nevertheless, mice present some important drawbacks: their inflammatory and immune reactions are far from those of human's, as is the physical structure of their lungs (in particular), which would be essential for the study of pulmonary diseases. There are many animals that have been tested as *in vivo* models for tuberculosis research such as mice, guinea pigs, and macaques [154, 113, 54] and in COVID-19 research such as mice, hamsters, and pigs [117, 139], but they are still far from human anatomy.

The Center of Comparative Medicine and Bioimage of Catalonia (CMCiB) is a center in the Can Ruti campus (Badalona) associated with the Institute for Health Science Research Germans Trias i Pujol (IGTP) where there are P3 security level laboratories able to carry out high security animal experiments with COVID-19 or TB, for example. In fact, the Tuberculosis Experimental Unit in IGTP is currently testing the use of mice and minipigs as *in vivo* models for tuberculosis research [12], with promising results. Recently, a mouse model was also used as COVID-19 animal model by IRSICAIXA, another institution on the same campus. Computational Biology and Complex Systems group (BIOCOM-SC) at the Universitat Politècnica de Catalunya of which I am a part collaborates with CMCiB where I work in the refinement and development of animal models: a minipig model to study tuberculosis and a mouse model to study stoke consequences.

4.2 Mathematical models

Mathematical models can be classified according to different criteria (e.g., empirical or mechanistic, spatially explicit or implicit, and continuous or discrete, among others) [81]. Among all these differences, one that we want to highlight is the difference between an empirical model and a mechanistic one. On the one hand, a mechanistic model, which is based on the underlying biological/physiological/physical processes, usually has as many features of the primary system as possible built into it as observations or data will allow. In contrast, an empirical model describes the system without all these precise details of how things work at a lower level, being based on the empirical observation of the dynamics of the variables of interest. As for the mechanistic models, we need a lot of knowledge about the described processes. In contrast, the building and use of empirical models does not require this knowledge. Therefore, they are used in systems when basic knowledge is limited [163].

In this thesis, three mechanistic and two empirical models are presented. One important aspect (related, in fact, to the difference between mechanistic and empiric models) is the strategy that is used to build them, which may be either top-down or bottom-up.

A top-down strategy would cover, for instance, the description of a certain system's dynamics by means of global differential equations. This strategy is useful and powerful when the heterogeneities of the system are not relevant to the problem addressed, and if the addressed variables fit the continuum hypothesis.

An example of top-down model is the **mathematical fitting**. The easiest model, it consists in adjusting a curve to a set of data. If done correctly, these models can be very powerful and used to predict or understand some basic behaviours. In fact, it is not necessary that a model be very complicated to be useful. Some examples of mathematical fittings are: exponential fitting to model cell growth [116, 14], a logistic model to fit lesion growth [140], and a Gompertz model used to estimate population survival [72] and to estimate bacterium counts in a controlled environment [86].

Another example is the **reaction-diffusion model**, a mathematical model used to describe several physical phenomena. It is mostly used in chemistry to simulate the evolution of local chemical reactions in which the substances are transformed into each other, and the diffusion which

causes the substances to spread out over a surface in space, for example to understand the kinetics of a reaction [173]. However, the system can also describe dynamic processes of a non-chemical nature such as, for instance, biological pattern formation [93].

Compartmental models are a particular representation of reaction-diffusion models. Usually, they do not have explicit space. They are a kind of model where different states are set and there are certain rates (or rules) of transitions between them. It can be applied to molecules, cells, individuals, and more. Like all models, they can be used in many different ways, from the study of pulmonary nitric oxide exchange dynamics [166] to the study the evolution of an epidemic [107, 3], to cite a few.

SIR or SEIR models are a particular case of compartmental models. Their name refers to the susceptible - (exposed) - infected - recovered dynamics of infectious diseases. In fact, they are particular compartmental model of 3/4 states. They have been classically used to model the evolution of an epidemic [179]. For example, a SEIR model was built to study the influenza pandemic in Japan [151].

The bottom-up strategy requires the modelling of the parts of the system and of the interactions between them in order to reproduce the whole-system dynamics. This is an interesting approach whenever emergent behaviors are expected to arise. The most commonly used examples of bottom-up models are **Individual-based** and **Agent-based models** (IbM and AbM):

Grimm [76] defines agent-based models as "simulation models that treat individuals as unique and discrete entities which have at least one property in addition to age that changes during the life cycle." They have been widely used in ecology and in microbiology [94], as well as in other disciplines like the social sciences. The building of an AbM requires the modelling of the behavior of an agent in a certain environment and its implementation in computer code. The simulation of a set of individuals acting simultaneously allows the observation of emerging patterns. This individual interacts with others under a certain set of rules that drives the evolution of each individual. In fact, depending on the subject of study there can be very different observed dynamics and rules. AbM are widely used on different scales, for example, to study the evolution of a TB lesion [29] or to study HIV transmission in a certain country [73].

4.2.1 Mathematical models in TB

Mathematical models in tuberculosis research have been mainly incorporated at the epidemiological level [30]. Most of them are compartment models, although some AbM have been incorporated in recent years [80]. At lower levels, mechanistic mathematical models have not been widely used, although there are some interesting examples. For instance, Bru and Cardona [22] presented a cellular automaton (close to an AbM) that aimed to describe granuloma (lesion) formation in mouse lungs. A widely used model is the Gran-Sim [107, 108] an AbM at cellular level (with diffusion to simulate space) that is able to reproduce some kinds of TB lesions.

Prats et al. [140] built an AbM denoted "bubble model" for tackling the lesions dynamics in TB infected mice. It considered the logistic growth of active TB lesions in a homogeneous space, as well as the merging of neighboring lesions and the appearance of new daughter lesions. The results of the simulations fitted available experimental mouse animal model data. Thus, the model allowed the quantification of the theoretical assumptions behind it, and their role in the global dynamics.

4.2.2 Mathematical models in COVID-19

For obvious reasons, there were no mathematical models of COVID-19 at the start of this thesis. Nevertheless, we may identify some models that could be applied to COVID-19 if modified properly. For example, models based on SIR or SEIR were able to reproduce the Ebola epidemic in 1995 [99], the SARS epidemic in 2002 [122] and the Ebola epidemic in 2014 [5]. Also, other models quantify the relation between different cells and viruses in the host [18] or they are used to forecast influenza for example using Google flu trends [62]. Empiric models such as logistic model and Gompertz model were also used to model and predict Ebola and SARS epidemics evolution [135, 24].

Initially, most COVID-19 models failed [87] for three reasons: (1) the lack of preliminary knowledge about the natural history of COVID-19 and its infection process; (2) problems with the data of diagnosed cases and deaths, which underestimated the real magnitude of the pandemic; and (3) the point that the global dynamics did not emerge from the natural dynamics of the infection, but was driven by external interventions such as NPI. Nevertheless, there are a lot of models that have been built

during this period with some interesting results. For example, in China, artificial intelligence was used to obtain predictions about the evolution of COVID-19 cases [83]. Another example is the iterative model that Perc et al. built to predict the evolution of COVID-19 epidemic in different countries using the number of cases in the previous days [111]. Compartmental models are also used to forecast COVID-19 evolution [152]. An interesting review that contain and compare some of this models can be found in: [147].

5 Context and goals

5.1 Context

As explained in the preface, this thesis started with two years of background in biophysics and tuberculosis modelling and some previous work [36, 34]. The main idea was to develop a thesis to understand tuberculosis infection dynamics by developing different scale models to study the impact of tuberculosis from the alveolus level to the community level. These models were to be fed with different animal model experimental data: minipigs, macaques, goats and, finally, adapting it to human data.

As can be seen in the time line of Figure 9 the main core of TB objectives was covered between 2017 and 2019. The coronavirus disease pandemic was like an earthquake for this thesis. With emergence of the pandemic, in February 2020, the Computational Biology and Complex Systems research group changed its study topic to COVID-19 to try to help in the complicated situation that we were experiencing. We have tried to be useful and use our knowledge in the resulting problems and challenges that the COVID-19 pandemic caused.

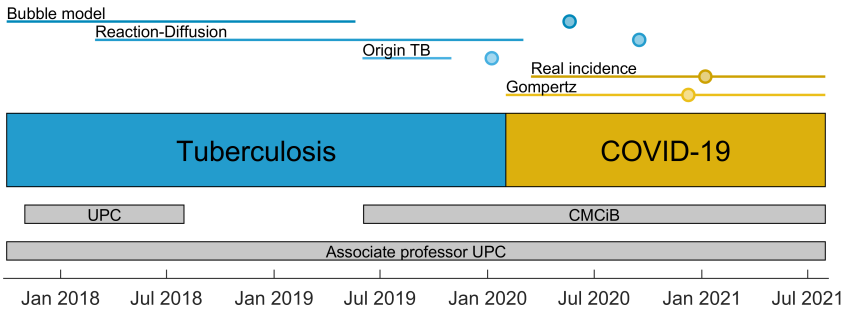

Figura 9: Thesis time line. Tuberculosis study period and coronavirus disease 2019 study period are marked. The working period in the different articles is marked with a solid line; each colour represents a different article, from top to bottom: modelling the dynamics of tuberculosis lesions in a virtual lung (Chapter 7), a reaction-diffusion model to understand granulomas formation (Chapter 6), origin of tuberculosis in the paleolithic (Chapter 8), robust estimation of diagnostic rate (Chapter 9) and empirical model for short-time COVID-19 prediction (Chapter 10). Article publication dates are marked with circles.

We have collaborated with the main public health institutions (Agència de Qualitat i Avaluació Sanitàries de Catalunya and Institut Català de la Salut) delivering more than 200 reports addressing different issues (predictions, future scenarios, variant substitution, and vaccination strategy, among others). We have also delivered periodic reports to DG-CONNECT European Commission department. On 14 June 2021 report 243 was delivered. In these reports we analysed many different topics related with COVID-19 consequences, forecasts, and analysis. The first report was delivered on 14 March 2020. Since that day the team has evolved and nowadays reports are signed by 9 investigators assisted by 5 students. During these months, more than 30 different investigators and more than 20 students have contributed to these reports. We have also collaborated with other national governments and institutions throughout the world.

During the development of this thesis I have been an associate professor in the Physics Department of the Universitat Politècnica de Catalunya. Between November 2017 and July 2018 I was research assistant investigator under the supervision of Dr. Sergio Alonso at the Universitat Politècnica de Catalunya. Since 3 June 2019 I have been the *in silico* technician in the Comparative Medicine and Bioimage Center of Catalonia which is an adjunct center of Germans Trias i Pujol Research Institute. CMCiB is a center focused on the 3Rs principle (Reduction, Refinement, and Replacement) in animal experimentation. Dr. Clara Prats and myself were in charge of the center branch dedicated to enhance the use of computational models. This thesis is part of the work carried out there.

5.2 Main objectives and goals

The aim of this thesis is to improve understanding of tuberculosis and coronavirus disease 2019 and the processes related to them. The specific objectives of this thesis are:

1. To build a model that simulates *Mycobacterium tuberculosis* infection in an alveolus (1D), determining the possible outcomes and the parameters that cause them.

2. To build a secondary lobe (3D) model that simulates the encapsulation process if a lesion is not contained inside the alveolus, determining the parameters which affect to their size.

3. To build a model to falsify dynamic tuberculosis hypothesis adjusting to a set of experimental observations of minipig latent tuberculosis infection.

4. To explore the transition between latent tuberculosis infection and active tuberculosis disease, determining the model parameters that cause it and their analogous biological factors and processes.

5. To build a SEIR model that simulates human-*Mtb* coexistence and takes into account the female protection and resistance, determining the role of this factor in this large coexistence.

6. To build a competitive model to simulate the substitution of "ancient" *Mtb* lineages by the so-called "modern" lineages, determining which factors propitiated this substitution.

7. To build a model to estimate real coronavirus disease 2019 incidence during the so-called first wave in Europe (March-May 2020) using reported COVID-19 mortality.

8. To determine the estimated active incidence of cases in studied countries and to compute the delays between case, detection and death in each country.

9. To build an empirical model that fits the evolution of cumulative and new cases of COVID-19 epidemic, adjusting the parameters to different countries and territories.

10. To use the empirical model to perform predictions calibrating the confidence intervals and obtaining the expected errors in a retrospective study for the countries that reported more than 1000 cases in May 2020.

Part II.

Articles

6 A reaction-diffusion model to understand granulomas formation inside secondary lobule during tuberculosis infection

6.1 Summary

The aim of this article is to create a model that simulates how TB infection is developed at the alveolus level when *Mtb* first arrives. The model also includes the diffusion to the nearest alveoli of the secondary lobe and how the lesion is encapsulated by fibroblast if the immune and inflammatory responses are not sufficient to contain *Mtb* progression.

Firstly, the 1D alveolus model is built to simulate the interaction between different cells and bacilli when it arrives at the alveolus. This interaction is characterized by 3 steps. (1) In the initial infection, bacilli are replicated because they can resist macrophages bactericidal mechanisms and use them to replicate inside. The successive replication processes cause macrophages lysis. (2) In the inflammatory response, new macrophages enter with neutrophiles in an attempt to stop the bacterial infection. Nevertheless, this has the oposite effect. Bacilli resist macrophages and neutrophil bactericidal mechanisms and use them to duplicate intracellularly (inside macrophages) and extracellularly (at neutrophils surface). (3) In the immune response, T cells arrive at the alveolus and activate macrophages, which (when activated) are able to eliminate bacilli. If the number of T cells and macrophages is enough compared to the bacilli they can eliminate the bacterial infection. These 3 steps are a simplification that allows us to simulate the infection of *Mtb* in an alvelous.

The model consists of 9 equations that account for the interaction between the different elements. Each transition is interpreted as a Poisson process to assure that the number of cells is always a non-negative integer. After running simulations with different parameter configurations, it is observed that, depending on the parameters set, the simulation can end in 3 possible states: i) proliferative lesion, in which immune response controls the infection and bacilli are eliminated, and after bacilli elimination inflammatory response is also stopped; ii) exudative lesion, in which inflammatory and immune response are not able to control infection and alveolus bacilli survive, so the alveolus is filled with cells because inflammatory response is not stopped; iii) bistability, in which both previous stages are possible, and we can compute the probability

of ending in each of the stages.

Parameters that determine each possible state are identified, the four most important being the duplication time of bacilli that increases the probability of exudative lesions and lifespan of T cells, activation rate of macrophages, and entering T cell flow that increases the probability of proliferative lesion.

Secondary lobe is considered to be a 50x50x50 grid of alveoli. In addition to the reactions that take place at the alveolus level, we include a diffusion term for each of the quantities that can move to near alveoli. At the secondary lobe boundary there is an infinite source of fibroblasts that follow macrophage gradients (chemotaxis).

With that secondary lobe model, it can be observed that exudative lesions are encapsulated as observed experimentally [70]. The final volume of lesions is determined by diffusion parameters. The faster fibroblast difuse, the smaller the lesion is. In contrast, the faster macrpohages and T cells diffuse, the bigger the lesion is. There are other parameters that are identified to help reduce lesion size. For instance, an important parameter is where the initial infection is situated with respect to the fibrobast: the closer it is to pulmonary membrane, the smaller the lesions formed.

Finally, we observe that lesion growth over time can be adjusted with a logistics fitting.

6.2 Main contributions

The main contributions of this article are:

- A model that reproduces the main steps of a Mtb infection in an alveolus is built.
- This model provides three possible outcoming dynamics: proliferative lesion, exudative lesion, and bistability. Parameters that determine the final state are identified and related to biological processes and factors.
- A 3D model is built as a group of alveoli with diffusion of cells and bacilli between them. Encapsulation is also incorporated with the presence of a fibroblast membrane that surrounds all secondary lobe spaces.

- Parameters that determine lesion size are identified and related to biological processes and factors. One of the most important parameters to determine a lesion's final size is the distance to the nearest pulmonary membrane.

6.3 Further work

Future steps include building a computational lung that can reproduce TB disease at the alveoli level that includes the bronchial tree structure and secondary lobe separation with pulmonary membranes. It would be of interest to reproduce minipig experimental data with this model [12].

The model could be refined if some intermediate steps that were simplified in the processes of macrophage clearance or immune activation were taken into account. In order to study the post-lesion formation scenario, it must be considered that when encapsulation is completed foamy macrophages cannot travel outside the lesion.

In the whole lung computational model, many parameters that do not have relevance in a single secondary lobe model should be taken into account such as ventilation, circulatory system, and different lobe sizes and shapes, among others. This may change some parameters and produce different lesion outcomes.

Lesions much larger than a single secondary lobe are produced in active tuberculosis. In order to model the break of the pulmonary membrane, some other mechanical processes must be considered.

6.4 Article

The full article can be read online at:
Català M, Prats C, López D, Cardona PJ, Alonso S. A reaction-diffusion model to understand granulomas formation inside secondary lobule during tuberculosis infection. *PLOS One*. 2020 Sep 16;15(9):e0239289. `https://doi.org/10.1371/journal.pone.0239289`

Supplementary video of the lesion encapsulation process: `https://doi.org/10.1371/journal.pone.0239289.s001`

PLOS ONE

RESEARCH ARTICLE

A reaction-diffusion model to understand granulomas formation inside secondary lobule during tuberculosis infection

Martí Català[1,2], **Clara Prats**[1,2]*, **Daniel López**[1], **Pere-Joan Cardona**[2,3,4], **Sergio Alonso**[1]

1 Department of Physics, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain, **2** Comparative Medicine and Bioimage Centre of Catalonia (CMCiB), Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol. Badalona, Catalonia, Spain, **3** Experimental Tuberculosis Unit (UTE), Fundació Institut Germans Trias i Pujol (IGTP), Universitat Autònoma de Barcelona (UAB), Badalona, Catalonia, Spain, **4** Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain

* clara.prats@upc.edu

Citation: Català M, Prats C, López D, Cardona P-J, Alonso S (2020) A reaction-diffusion model to understand granulomas formation inside secondary lobule during tuberculosis infection. PLoS ONE 15(9): e0239289. https://doi.org/10.1371/journal.pone.0239289

Editor: Frederick Quinn, The University of Georgia, UNITED STATES

Received: May 12, 2020

Accepted: September 3, 2020

Published: September 16, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: CP, PJC and MC received funding from La Caixa Foundation (ID 100010434), under agreement LCF/PR/GN17/50300003; PJC received funding from Agència de Gesto d'Ajuts Universitaris i de Recerca AGAUR), Grup Unitat de Tuberculosi Experimental, 2017-SGR-500; CP, DL, SA, MC received funding from Ministerio de

Abstract

Mycobacterium tuberculosis (*Mtb*) is the causative agent for tuberculosis, the most extended infectious disease around the world. When Mtb enters inside the pulmonary alveolus it is rapidly phagocytosed by the alveolar macrophage. Although this controls the majority of inhaled microorganisms, in this case, Mtb survives inside the macrophage and multiplies. A posterior chemokine and cytokine cascade generated by the irruption of monocytes, neutrophils and posteriorly, by T-cells, does not necessarily stop the growth of the granuloma. Interestingly, the encapsulation process built by fibroblasts is able to surround the lesion and stop its growing. The success of this last process determines if the host enters in an asymptomatic latent state or continues into a life-threatening and infective active tuberculosis disease (TB). Understanding such dichotomic process is challenging, and computational modeling can bring new ideas. Thus, we have modeled the different stages of the infection, first in a single alveolus (a sac with a radius of 0.15 millimeters) and, second, inside a secondary lobule (a compartment of the lungs of around 3 cm^3). We have employed stochastic reaction-diffusion equations to model the interactions among the cells and the diffusive transport to neighboring alveolus. The whole set of equations have successfully described the encapsulation process and determine that the size of the lesions depends on its position on the secondary lobule. We conclude that size and shape of the secondary lobule are the relevant variables to control the lesions, and, therefore, to avoid the evolution towards TB development. As lesions appear near to interlobular connective tissue they are easily controlled and their growth is drastically stopped, in this sense secondary lobules with a more flattened shape could control better the lesion.

Introduction

Tuberculosis (TB) is an infectious disease that on 2017 killed 1.6 million people [1]. The same year, nearly 10 million people developed the disease. Despite our best efforts, *Mycobacterium*

Ciencia, Innovación y Universidades and FEDER, with the project PGC2018-095456-B-I00.

Competing interests: The authors have declared that no competing interests exist.

tuberculosis (*Mtb*) remains the bacteria able to cause the highest mortality by itself. Since World Health Organization declaration of TB public health emergency in 1994 [2], TB death rate has been reduced from 23% in 2000 to 16% nowadays [1]. In 2015, the TB End Strategy stated the goal of reducing its incidence by 50% in 2025 and 90% in 2035, as well as of reducing TB death to 75% in 2025 and 95% in 2035 [3]. It is estimated that between one quarter and one third of world population is infected with *Mtb* and that around a 10% of them will develop an active TB disease in the future years.

Tuberculosis natural history has an extraordinary complexity and, in fact, there are still too many unknowns [4]. Tuberculosis infection starts when an *Mtb* is phagocyted by an alveolar macrophage (AM) at a pulmonary alveolus. *Mtb* resists bactericidal mechanisms induced by AM and replicates inside [5]. This capacity is the biggest challenge of *Mtb* and the main reason of its high mortality in humans along history. Under proper conditions, *Mtb* replicates approximately once a day [6]. When the intracellular bacterial load overcomes AM maximum tolerability, macrophage necrosis is triggered and thereby bacilli enter the extracellular milieu. These bacilli are phagocyted by other AM, which also fail to control the bacillary growth and are likewise destroyed. This cycle is followed by a local inflammatory response and potentially ends once the specific immune response appears and controls it. This is the end of the progressive infection, which finally leaves an encapsulated TB lesion [6, 7]. Fibroblasts cells lead encapsulation with collagen, fibrin and other molecules. Fibroblasts cells are mainly located at pulmonary membranes like intralobular *septae*. In fact, this suggests that lungs structure may have an important role in TB infection dynamics, since initial infection distance to the nearest pulmonary membrane may determine the encapsulation capacity and, therefore, the final lesions size and bacillary load. Lungs are composed by around 2500 secondary lobules, and each lobule has an approximate volume of 3 cm^3.

There are two types of lesions: proliferative and exudative lesions. Proliferative lesions are mainly based in macrophages and lymphocytes, with intracellular bacilli inside the macrophages and minimal necrosis, controled through time; exudative lesions are mainly based on necrotic tissue caused by neutrophilic infiltration where a lot of extracellular bacilli are accumulated [8–10].

If the response of the host is adequate, an *Mtb* infection can be completely cleared from the organism [11] or it can enter an asymptomatic latent state where the host is infected but not sick and cannot infect other people, corresponding to latent tuberculosis infection (LTBI). Nevertheless, if the immune and the inflammatory responses are not correctly balanced, the host can develop a more compromised situation, the active tuberculosis disease (ATB).

Systems biology and computational models are great tools for increasing TB understanding [12]. Last years, several TB models have been built for a better understanding of different processes related with TB natural history [4]. In particular, GranSim [13] is a hybrid model that, under several modifications, is able to reproduce granulomas formation [14], encapsulation [15] and study drug resistance [16], among others. Bubble model [17, 18] is an Agent-based Model that reproduces the formation and growth of lesions, as well as the coalescence of neighboring lesions and their spreading by means of bronchial endogenous reinfection processes. Other mathematical models study different states in macrophage-bacilli competitivity [19].

In this work we develop a mathematical model to reproduce *Mtb* infection dynamics in an alveolus by means of Stochastic Differential Equations (SDEs). It is then generalized to a three-dimensional stochastic reaction-diffusion model that incorporates the space to reproduce *Mtb* infection, lesions formation and encapsulation in a secondary lobule. Finally, we identify the factors that determine lesions size and encapsulation time.

Materials and methods

Model structure

Two different approaches are considered: the modeling of a single alveolus and the modeling of a secondary lobule. The difference between both models is the dimensionality of the space where the simulation takes place. Single alveolus system is a non dimensional model, therefore modeled by ordinary differential equations, and the secondary lobule is a three dimensional system modeled by partial differential equations.

Next we explain the interactions among the different cells and elements of the model. We consider 5 types of cells in this model: bacilli (b), macrophages (m), neutrophils (n), T cells (T) and fibroblasts (f). Bacilli are distinguished between intracellular (b_I), i.e. bacilli inside a macrophage, and extracellular (b_E), i.e. bacilli in the extracellular milieu. Accordingly, macrophages can be uninfected (m_U), i.e. they do not have bacilli inside, infected (m_I), i.e. they present bacilli inside but are not able to eliminate them, or activated-foamy (m_{AF}), i.e. they present bacilli inside, are able to eliminate extracellular bacilli (activated) and drain intracellular ones (foamy). Volume occupied by death elements (V_O) and local inflammatory response (s) are also considered. Therefore, we consider the evolution of 8 types of cells and 2 additional processes. We take into account 16 processes, which lead to 10 reaction-diffusion equations, one for each element. These processes are described below. Note that the numbers in parentheses relate to the description with the corresponding terms in the model equations and with the processes depicted in Fig 1.

While the initial infection is closely related with the dynamics inside of an alveolus, the final part of the process, the encapsulation of the granuloma, takes into consideration the structure of the secondary lobule. For the first modeling appoach based on a single alvoulus, we study the effect of inflammatory and immune responses at the level of the alveolus, while for the second approach such responses are considered at a larger scale, the secondary lobule. For both approaches, we divide the whole evolution in four differential modules:

Initial infection. *Mtb* infection starts when an AM phagocytes a bacillus in the alveolar surface (3, *phagocytosis by non-infected macrophages*). In this process, the extracellular bacilli become intracellular and non-infected macrophage becomes infected. *Mtb* resist bactericidal mechanisms induced by AM and replicates inside it (1, *intracellular bacilli growth*), as they use infected macrophages as sustenance. When intracellular bacillary load overcomes macrophage maximum tolerability (α), macrophage necrosis is triggered and thereby bacilli enter the extracellular milieu (5, *macrophages' lysis*). Intracellular bacilli released become extracellular and the infected macrophage is eliminated. Then, other AM from closer alveoli enter the infected alveolus and the bacilli growth-macrophage lysis process occurs again. Infected macrophages also phagocyte extracellular bacilli (4, *phagocytosis by infected macrophages*) so that these extracellular bacilli also become intracellular. An explanatory schematic of this module is shown in Fig 1A.

Inflammatory response entrance. Macrophages death and bacillary load's increase trigger inflammatory response (7, *inflammatory response activation*). The variable s is increased from its initial value, 0, up to 1 after 3-5 days. This value is an approximation based on experimental observations [20]. When $s = 1$, inflammatory response is fully active. This increase in s leads to an entrance of neutrophils and uninfected macrophages in the alveolus (6, *inflammatory entrance*). The presence of neutrophils may be used by extracellular bacilli as a sustenance to grow (2, *extracellular bacilli growth*). There is a limit of how many extracellular bacilli can grow on a neutrophil (δ). If bacillary load decreases and infected macrophages disappear, inflammatory response is inhibited (8, *inflammatory response inhibition*). Macrophages and neutrophils cascade is unable to contain the infection, in fact, bacillary load grows exponentially. An explanatory scheme of this module is shown in Fig 1B.

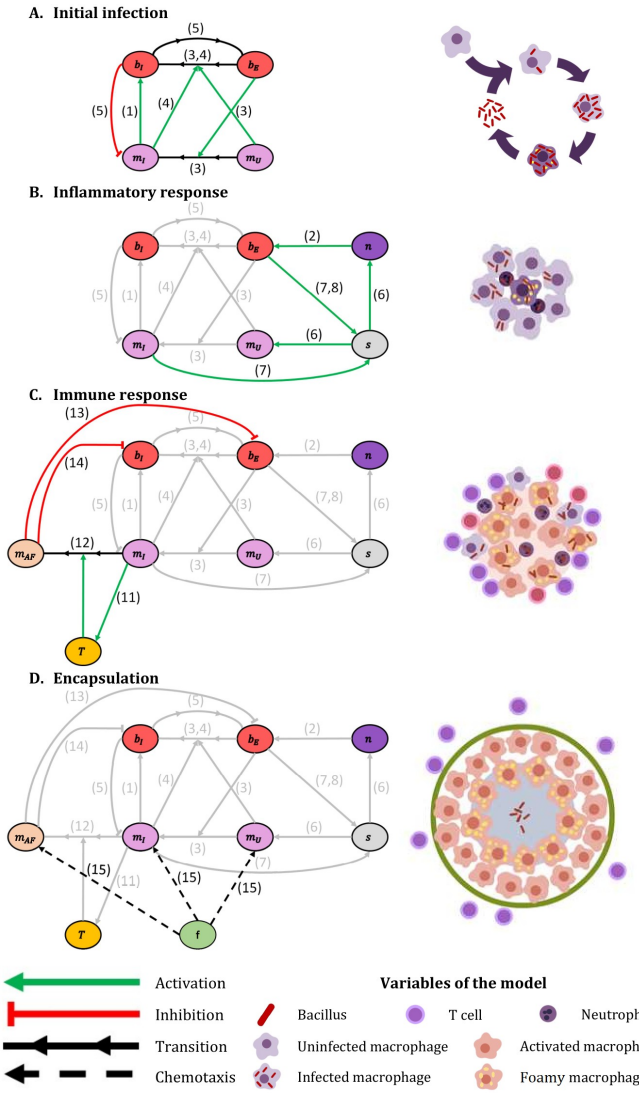

Fig 1. Four principal model's modules. Interaction between different model elements. (A) Initial infection: interaction between bacilli and macrophages. (B) Inflammatory response: triggered by the presence of extracellular bacilli. (C) Immune response: infected macrophages trigger the mechanisms that activate this response. (D) Encapsulation: led by fibroblasts, it takes place when a lesion makes contact with a pulmonary membrane. Green arrows: activation; red arrows: inhibition; black arrows: transition; dotted arrows: chemo-taxis. Numbers identify specific processes: 1 intracellular bacilli growth, 2 extracellular bacilli growth, 3 phagocytosis by non-infected macrophages, 4 phagocytosis by infected macrophages, 5 macrophages' lysis, 6 inflammatory response entrance, 7 inflammatory response activation, 8 inflammatory response inhibition, 11 immune response entrance, 12 macrophages activation, 13 extracellular bacilli elimination, 14 intracellular bacilli drainage and 15 fibroblasts chemo-taxis. See model's details in Eqs (1)–(10). Part of this figure is adapted from [6].

https://doi.org/10.1371/journal.pone.0239289.g001

Immune response. Some infected macrophages trigger dendritic cells that travel to Lymph node to trigger immune response. Immune response consists of a T cells flow to the affected alveoli (11, *immune response entrance*). Infected macrophages are activated by contact with T cells (12, *macrophages activation*). This activated macrophages are able to phagocyte extracellular bacilli and eliminate them (13, *extracellular bacilli elimination*). Activated macrophages have a short lifespan. When they are not able to eliminate more bacilli they convert to foamy macrophages [21–23]. These foamy macrophages leave the alveolus by draining the intracellular bacilli inside them (14, *intracellular bacilli drainage*). All the elements have a certain lifespan; when they die, they leave a corpse that occupies a fraction of the space (9, *death rate*). A scheme of this module is shown in Fig 1C. Even when the living-time of the neutrophils is limited (between 5 and 135 hours [24]), as they made neutrophilic extracellular nets (NETs) [25], and become part of the necrotic tissue, we have extended significantly the time of degradation. An explanatory scheme of this module is shown in Fig 1C.

Encapsulation. During all these processes all elements diffuse, therefore occupying surrounding alveoli (16, *diffusion*). All elements diffuse by their own except intracellular bacilli, whose diffusion is conditioned by infected macrophages that contain them. The initial infection forms a lesion due to cell accumulation. When lesions size is big enough to get in contact with pulmonary lobule membranes (interlobular connective tissue), encapsulation process is triggered. Encapsulation is led by fibroblasts cells that chemo-tact macrophages gradient (15, *fibroblast chemo-taxis*). Initially, fibroblasts cells are located at pulmonary lobules membranes. An explanatory scheme of this module is shown in Fig 1D.

In Fig 1, a summary of the relations between the 10 variables of the model are shown, together with the different interactions involved, organized in the four modules previously mentioned. The corresponding 10 reaction-diffusion equations of the model are:

$$\frac{\mathrm{d}}{\mathrm{d}t}b_I = \underbrace{\mu b_I\left(1-\left(\frac{b_I}{\alpha m_I}\right)^3\right)}_{(1)} + \underbrace{\gamma g m_U b_E}_{(3)} + \underbrace{\gamma g m_I b_E\left(1-\left(\frac{b_I}{\alpha m_I}\right)^3\right)}_{(4)} - \underbrace{\alpha(\beta b_I+(1-\beta\alpha)m_I)}_{(5)} - \underbrace{\lambda_b b_I}_{(9)} - \underbrace{\frac{\alpha}{\tau_m}m_{AF}}_{(14)} + \underbrace{\nabla\left(gD_m\frac{b_I}{m_I}\nabla m_I\right)}_{(16)}, \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}b_E = \underbrace{\mu b_E\left(1-\left(\frac{b_E}{\delta n}\right)^3\right)}_{(2)} - \underbrace{\gamma g m_U b_E}_{(3)} - \underbrace{\gamma g m_I b_E\left(1-\left(\frac{b_I}{\alpha m_I}\right)^3\right)}_{(4)} + \underbrace{\alpha(\beta b_I+(1-\beta\alpha)m_I)}_{(5)} - \underbrace{\lambda_b b_E}_{(9)} - \underbrace{\nu m_{AF} g}_{(13)} + \underbrace{\nabla(gD_b\nabla b_E)}_{(16)}, \tag{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}m_U = -\underbrace{\gamma g m_U b_E}_{(3)} + \underbrace{k_1\left(\frac{b_E}{m}\right)sg}_{(6)} - \underbrace{\lambda_m m_U}_{(9)} + \underbrace{\nabla(gD_m\nabla m_U)}_{(16)}, \tag{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}m_I = \underbrace{\gamma g m_U b_E}_{(3)} - \underbrace{(\beta b_I+(1-\beta\alpha)m_I)}_{(5)} - \underbrace{\xi m_I T g}_{(12)} + \underbrace{\nabla(gD_m\nabla m_I)}_{(16)}, \tag{4}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}m_{AF} = \underbrace{\xi m_I T g}_{(12)} - \underbrace{\frac{1}{\tau_m} m_{AF}}_{(14)} + \underbrace{\nabla(g D_m \nabla m_{AF})}_{(16)}, \tag{5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}n = \underbrace{k_2\left(\frac{b_E}{m}\right) s g}_{(6)} - \underbrace{\frac{1}{t_n} n}_{(10)} + \underbrace{\nabla(g D_n \nabla n)}_{(16)}, \tag{6}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}T = -\underbrace{\lambda_T T}_{(9)} + \underbrace{\kappa m_I s g}_{(11)} + \underbrace{\nabla(g D_T \nabla T)}_{(16)}, \tag{7}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}f = \underbrace{\nabla(g\chi_f |\nabla g| \nabla f)}_{(15)}, \tag{8}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}V_O = \underbrace{(\beta b_I + (1-\beta\alpha)m_I)V_m}_{(5)} + \underbrace{\lambda_m m_U V_m}_{(9)} + \underbrace{\lambda_T T V_T}_{(9)} + \underbrace{\frac{1}{t_n} n V_n}_{(10)}, \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}s = \underbrace{\rho_U m_I b_E (1-s)}_{(7)} - \underbrace{\rho_D \frac{s}{b_E + 1}}_{(8)}; \tag{10}$$

where g is the unoccupied fraction of volume in each alveolus:

$$g = 1 - \frac{(m_U + m_I + m_{AF})V_m + n V_n + T V_T + V_O}{V_a}, \tag{11}$$

and $k_1(x)$ and $k_2(x)$ are the entrance rate of macrophages and neutrophils, respectively. They are quadratic functions that depend on extracellular bacilli—macrophages ratio. When this ratio is high enough, immune response is neutrophils-based (NBR) and most of the entering elements are neutrophils, otherwise, if macrophages are able to control extracellular bacilli, the immune response is macrophages-based (MBR) and most of the entering elements are macrophages. The threshold that differentiates both is half of macrophages tolerability ($\alpha/2$). The total amount of external agents flux (Y) is conserved:

$$k_1(x) \cdot V_m + k_2(x) \cdot V_n = \Upsilon, \tag{12}$$

$$k_1(x) = \frac{\Upsilon}{V_m} \frac{4\epsilon^+ x^2 + \epsilon^- \alpha^2}{4x^2 + \alpha^2}, \tag{13}$$

$$k_2(x) = \frac{\Upsilon}{V_n} \frac{4(1-\epsilon^+)x^2 + (1-\epsilon^-)\alpha^2}{4x^2 + \alpha^2}. \tag{14}$$

Macrophages and neutrophils entrance rate profile per day can be observed in Fig 2 for values of extracellular bacilli and total amount of macrophages ratio between 0 and 2α.

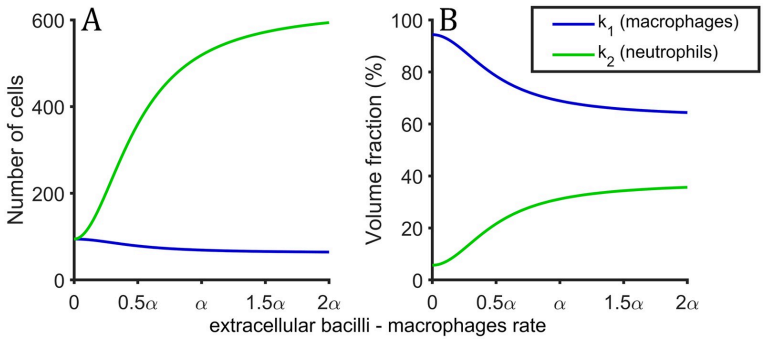

Fig 2. Inflammatory response profile. Profile of the inflammatory response as a function of extracellular bacilli and total amount of macrophages. (A) Number of cells that enter per day at the alveoli when inlfammatory response is fully activated. (B) Volume fraction of macrophages (blue) and neutrophils (green) that enter in alveoli due to inflammatory response.

https://doi.org/10.1371/journal.pone.0239289.g002

Note that the 3D diffusion equations part is marked in blue. If the blue parts of the equation are not considered, we recover the equations for a single alveolus model.

Parameters

This model depends on a total of 27 parameters that are obtained from bibliography or adjusted from experimental observations. All the parameters, their baseline value and corresponding references are shown in Table 1. Parameters Y, κ and ξ are adjusted to reproduce Bru & Cardona observations [26]. Parameter δ is adjusted from Marzo et al figures [20]. Parameter ρ is adjusted from Marzo et al observations [20]. Parameters β and γ are computed using simple Agent-Based Models (ABM).

Adjusting β. The parameter β is fitted by means of an ad-hoc ABM designed for studying the distribution of b_I among m_I and the resulting amount of macrophages' lyses. We develop a simple ABM where macrophages are the agents with only one property: the number of bacilli inside them. All other elements behave like the model described before. There is no diffusion, as just one alveolus is considered (single alveolus model). Macrophage lysis is triggered when its bacillary load is higher than macrophages maximum tolerability (α). Macrophages death rate due to bacilli growth is computed at each time step. As a result of the simulation, a function that relates the number of intracellular bacilli, number of infected macrophages and lysis rate due to intracellular bacilli growth can be fitted:

$$\text{lysis}(m_I, b_I) = \beta_1 b_I + \beta_2 m_I; \tag{15}$$

where β_i are the adjusted constants. Fitting goodness R^2 is computed and values over 0.8 are obtained for different parameters sets exploring diverse initial conditions. Using baseline values and a set of random initial conditions we obtain: $\beta_1 = 0.01816\ \text{day}^{-1}$, $\beta_2 = -0.1007\ \text{day}^{-1}$, $R^2 = 0.895$. Note that, as expected, the number of bacilli needed to kill one macrophage is α:

$$\beta_1 \cdot \alpha + \beta_2 \cdot 1 \approx 1. \tag{16}$$

Table 1. Parameters description and values. Parameter with source ABM is adjusted from Agent-Based Model, details are on the text. Parameter with source ABM* is fitted from ABM and corrections are implemented. Parameters with adj. source were adjusted to reproduce tuberculosis infection dynamics according to that source. Parameters D_b, D_m, D_n, D_T and χ_f are marked with $\star$ because they are not directly explored. These parameters are explored using D and DF parameters, related with Eq 24. Parameter D is explored between half and double its baseline value ($4.6{\cdot}10^{-9}$) and DF is explored between 0.5 and 4.

parameter	description	values	baseline value	units	explored values	source
μ	macrophages growing rate	0.69–1.04	0.693	day^{-1}	0.35-1.39	[6, 27]
α	macrophages maximum tolerability	32–64	60	-	30-120	[6]
β	bacilli kill macrophages	0.01–0.04	0.0182	day^{-1}	adj.	ABM
γ	phagocytation rate	$8{\cdot}10^{-4}$–8	0.05	day^{-1}	0.025-0.1	ABM*
δ	neutrophils capacity	10-40	30	-	15-60	adj. [20]
κ	T cells recuitment		0.8	day^{-1}	0.4-1.6	[19, 26]
ξ	macrophages activation		0.01	day^{-1}	0.005-0.02	[19, 26]
ν	activated killing rate		150	day^{-1}	75-300	adj. [6]
ρ_U	inflammatory activation		10^{-4}	day^{-1}	10^{-5}-10^{-3}	adj. [20]
ρ_D	inflammatory inhibition		1	day^{-1}	no	adj. [20]
Y	inflammatory flow		$5 \cdot 10^5$	$\mu m^3 day^{-1}$	$2 \cdot 10^5$-10^6	adj. [6]
ϵ^+	macrophages proportion in NBR		5/8	-	no	adj. [6]
ϵ^-	macrophages proportion in MBR		50/53	-	no	[28]
λ_b	bacilli death rate	$1\text{-}8 \cdot 10^{-3}$	10^{-2}	day^{-1}	0.005-0.02	[29]
λ_m	macrophages death rate		0.023	day^{-1}	0.01-0.05	[30]
τ_m	activated macrophages lifespan	3–10	5	day	2.5-10	[6, 31]
t_n	neutrophils degradation time		100	day	50-200	adj.
λ_T	T cells death rate	0.20–0.33	0.33	day^{-1}	0.16-0.66	[6]
V_m	macrophage volume		5000	μm^3	no	[32]
V_n	neutrophil volume		300	μm^3	no	[33]
V_T	T cell volume		200	μm^3	no	[34]
V_a	alveolus volume	$4 \cdot 10^6$–$6 \cdot 10^7$	10^7	μm^3	no	[35]
D_b	bacilli diffusion	$D_m/10$	$4.6 \cdot 10^{-10}$	$cm^2 s^{-1}$	$\star$	[31]
D_m	macrophages diffusion	10^{-5}–10^{-11}	$4.6 \cdot 10^{-9}$	$cm^2 s^{-1}$	$\star$	[31]
D_n	neutrophils diffusion	10^{-5}–10^{-11}	$4.6 \cdot 10^{-9}$	$cm^2 s^{-1}$	$\star$	[31]
D_T	T cells diffusion	10^{-5}–10^{-11}	$4.6 \cdot 10^{-9}$	$cm^2 s^{-1}$	$\star$	[31]
χ_f	fibroblasts chemotaxis coefficient		$9.2 \cdot 10^{-9}$	$cm^2 s^{-1}$	$\star$	adj. [6]

https://doi.org/10.1371/journal.pone.0239289.t001

Then, imposing Eq (16) and redefining β_1 as β we obtain a new expression for the lysis function:

$$\text{lysis}(m_I, b_I) = \beta b_I + (1 - \beta\alpha) m_I. \tag{17}$$

Using baseline values we obtain $\beta = 0.0182$ day^{-1} and $R^2 = 0.884$. The value of β is calculated at the start of every simulation and depend on the parameters set used. Parameters α and μ are the parameters that mostly determine β value.

Adjusting γ. The parameter γ is also estimated using another ad-hoc ABM that simulates a simple macrophages-bacilli interaction in an alveolus. We consider a 2D system where bacilli are considered fixed rectangles of width 0.6 μm and length 4 μm [36] and macrophages are mobile circles with radius $r_m = 10.6$ μm [32]. Simulations take place in a L × L square with periodic boundary conditions which represent alveolar surface, L = 532 μm [35]. Macrophages diffuse moving at constant velocity but with a changing direction. Using Fürth formula [37] a

diffusive movement can be characterized by the following equations:

$$\frac{\mathrm{d}}{\mathrm{dt}}\vec{x}(t) = \vec{v}(t), \tag{18}$$

$$\vec{v}(t) = v_0 \cdot (\cos\theta(t),\ \sin\theta(t)), \tag{19}$$

$$\frac{\mathrm{d}}{\mathrm{dt}}\theta(t) = \sigma \cdot \eta(t); \tag{20}$$

where v_0 is a constant velocity that depends on diffusive coefficient, D_m, and persistence time, τ_p, that for macrophages can be approximated as $\tau_p = 2000$ s [38]:

$$v_0 = \sqrt{\frac{2D_m}{\tau_p}}; \tag{21}$$

and $\eta(t)$ is white noise of amplitude σ that depends on persistence and time step, dt.

$$\sigma = \sqrt{\frac{2}{\tau_p \mathrm{dt}}}. \tag{22}$$

At each time step, macrophages move and phagocyte bacilli at a distance smaller than its radius, (r_m). These bacilli are eliminated from the system. Every time that a macrophage phagocytes a bacillus, macrophage is stopped during $t_{phag} = 10$ minutes [27]. At each time step, the number of bacilli, macrophages and phagocytations that take place are computed. Different random initial conditions are explored. As a result, the phagocyte ratio can be approximated as:

$$\mathrm{phagocyte}(m, b_E) = \gamma m b_E. \tag{23}$$

Different sets of parameters adjust to different γ values. R^2 is computed being always higher than 0.90. Using baseline values $\gamma = 0.56$ day^{-1} and $R^2 = 0.9882$. This simplified model give us a range of values for γ, $\gamma_{min} = 8 \cdot 10^{-4}$ day^{-1} ($D_m = 10^{-11}$ cm^2 s^{-1}) and $\gamma_{max} = 7.6$ day^{-1} ($D_m = 10^{-5}$ cm^2 s^{-1}). Due to biological factors as alveolus saturation, long range interactions and occupation factor we considered γ smaller and its baseline value was set to $\gamma = 0.05$ day^{-1}.

Implementation

The model was implemented in MATLAB using Poisson τ-leap method to integrate differential equations [39], $\tau = 0.1$ days for the one alveolus model and $\tau = 0.001$ days for the secondary lobule model. Diffusion was computed with finite differences, $\Delta x = 0.3$ mm. Transition probabilities were determined by the differential equation and the success or not was determined using Poisson random numbers to assure that all quantities are integers. Poisson random numbers are also used in finite differences.

For the secondary lobule model (3D simulations) we consider a cubic secondary lobule of 3.4 cm^3 (1.5 × 1.5 × 1.5 cm). Secondary lobule is formed by 125000 (50 × 50 × 50) alveoli, cubes of 0.3 mm side [35]. Reaction equations take place in each alveoli and diffusion is considered with surrounding alveoli. The cube is surrounded by a thin membrane (1 alveolus thickness) where fibroblasts concentration is maximum. Membrane fibroblasts are considered inexhaustible (boundaries are always full of fibroblasts).

Results

Model of a single alveolus

Two final states are observed: Proliferative and exudative scenarios. Single alveolus model results are computed eliminating diffusive (blue) parts of Eqs (1)–(10). Exploring different set of parameters and diverse initial conditions we see that there are two possible final states as it can be seen in Fig 3. (1) Proliferative scenario: bacilli population is eliminated by a flux of macrophages that are activated by T cells. Temporally, part of the alveolus is filled by macrophages, however, finally the number of macrophages returns to 1. Some of the space remains occupied by dead cells. (2) Exudative scenario: bacilli persevere and are able to kill all macrophages that try to contain the infection. Most of the present cells in this scenario are neutrophils thus favouring extracellular growth of bacilli. At the end, all bacilli are extracellular due to macrophages lysis and alevolus is occupied by dead cells of macrophages, neutrophils and T cells.

Under certain values of the parameters, bistability of both states is observed and the fluctuations determine the final state. According to these different simulation possibilities we can define three zones of parameters: Proliferative zone where all simulations end in a proliferative scenario, bistable region where a simulation can end in exudative or proliferative scenario depending on the particular realization, and exudative zone where all simulations end in exudative scenario. For each set of parameters we can compute the probability to end in exudative state, p_{exu}, or proliferative scenario, p_{pro}. Note that $p_{exu} = 1 - p_{pro}$.

Activation rate of macrophages determines transition between proliferative and exudative scenarios. In Fig 3, a sample of simulations with three different set of parameters are shown, one set for each defined zone (proliferative, bistable and exudative). The parameter used to explore different parameters zone is macrophage activation rate (ξ). Different values used were: proliferative ($\xi = 0.5$ day^{-1}), bistable ($\xi = 0.15$ day^{-1}) and exudative ($\xi = 0.01$ day^{-1}). For the set of parameters of the bistable zone, the probability to end in exudative scenario is $p_{exu} = 0.22$. This probability is computed using the results of 10000 simulations till day 200.

Initial behaviour is very similar for all sets of parameters, inflammatory response is triggered by first macrophages lysis that takes places 6.4 ± 2.3 days after initial infection and its fully awaken around day 16th. Firsts T cells arrive at the alveolus at day 14.2 (1 T cell), and around day 17th there are around 25. At day 16.7 ± 2.8, the first activated macrophage is observed, after 19th day more than 10 macrophages are active.

Transition to exudative scenario depends on bacilli growth and T cells and macrophage activity. We perform a sensitivity analysis to determine the input parameters effect in outcome variables. We employ the methodology described in [40]. Parameters are sampled using a Latin Hypercube Sample technique [41] between the values specified in Table 1. We explore a total of 14 parameters, which are those that do not have a well defined value. In most of the cases, the exploration ranges are between half and double of their baseline value, but in some parameters of higher interest exploration ranges are extended. We use 1500 points to perform the exploration. Partial Rank Correlation Coefficient (PRCC) is computed between input parameters and outcome variables along time. In Fig 4 we show the correlation between six parameters (μ, α, δ, γ, Y and ρ) and the outcomes intracellular bacilli (b_I), extracellular bacilli (b_E), and total macrophages (m). Parameters κ, ν, ξ, λ_b, λ_m, λ_T, τ_m and t_n are not shown in Fig 4 because they do not present correlations bigger than 0.2 with these outcome variables. It can be seen that as expected parameter μ has a positive correlation with the number of bacilli, long term correlations decrease because alveolus is filled. Initially μ reduces macrophages quantity due to bacilli growing inside, but later this effect triggers inflammatory response and macrophages numbers increase. Parameter α contributes to increase bacilli quantities because more

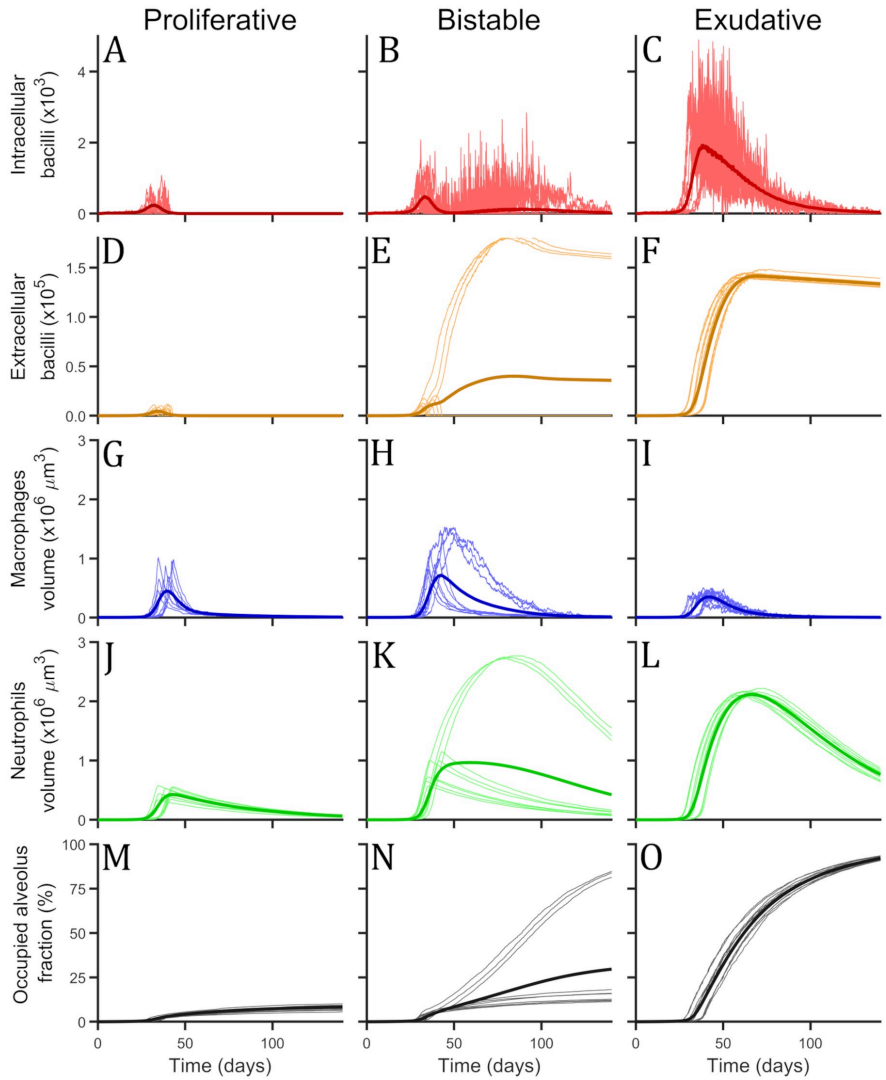

Fig 3. Single alveolus model simulations with three different set of parameters. (A, D, G, J, M) Proliferative set, (B, E, H, K, N) bistable set and (C, F, I, L, O) exudative set. For each set of parameters there can be seen the total amount of intracellular bacilli, in red (A, B, C), the total amount of extracellular bacilli, in orange (D, E, F), the total volume occupied by macrophages, in blue (G, H, I), the total volume occupied by neutrophils, in green (J, K, L) and the occupied alveolus volume fraction, in black (M, N, O). Thin and bright lines represent single simulations, 10 single simulations are shown as example for each set. Thick and dark lines are the mean value of 10000 simulations. Simulations were 200 days long, but only the first 140 are shown.

https://doi.org/10.1371/journal.pone.0239289.g003

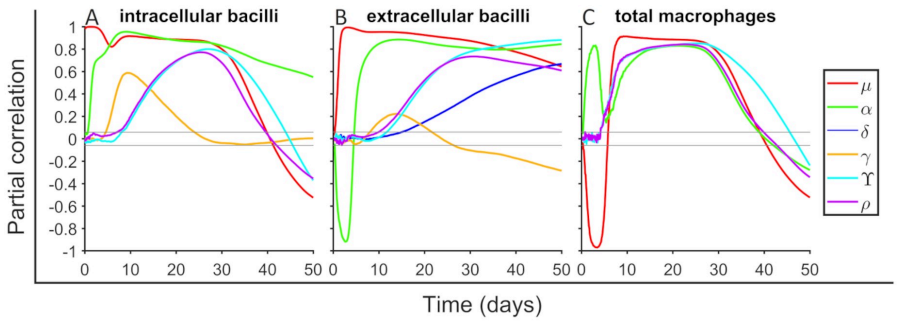

Fig 4. Sensitivity analysis for intracellular bacilli, extracellular bacilli and total macrophages. Partial Rank Correlation Coefficient (PRCC) is computed between a total of six input parameters (see legend) and outcome variables. A total of 1500 parameters sets are used for exploring the parameter space defined in Table 1 using Latin Hypercube Sample technique. A total of 14 parameters are explored, the ones that are not shown it is because their PRCC values is smaller than 0.2. Gray horizontal lines are marked at ±0.043 to mark significance threshold for PRCC values using a significance level of 0.05. PRCC are computed in each time step (dt = 0.1 day). For each set of parameters a total of 10000 simulations are done.

https://doi.org/10.1371/journal.pone.0239289.g004

bacilli can duplicate inside a macrophage, these also trigger inflammatory response and more macrophages enter to the alveoli. Parameter δ increases the number of extracellular bacilli because more bacilli can duplicate over a neutrophil. Parameters Y and ρ have a very similar effect because both are related with inflammatory response (Y increases the entering flux and ρ activates faster the inflammatory response). This accelerates the process and more bacilli and macrophages are seen before. Parameter γ is important to determine how fast are bacilli engulfed and its effect is mainly seen after first macrophage lysis.

As seen in Fig 3, the set of parameters determines final state probabilities. We compute PRCC between 14 input parameters used in sensitivity analysis and the probability to end in exudative scenario. In Fig 5 PRCC final values are shown. Increasing T cells death rate (λ_T) and macrophages growing rate (μ) or reducing macrophages activation (ξ), T cells recuitment (κ), activated macrophages killing rate (ν) and activated macrophages lifespan (τ_m) increase the probability to end in exudative scenario. Therefore, parameter ξ is not the unique parameter that determines the zone where a desired set of parameters belong.

Model of secondary lobule

Fibroblast encapsulation stops the expansion of the lesion in the model. Initially all alveoli present one uninfected macrophages except a certain alveolus that has an infected macrophage with a bacillus inside. The resulting evolution of the infection can be observed in Fig 6. We show the total bacilli amount (b), total macrophages amount (m) and fibroblasts concentration (f). In supplementary material there is a video with lesions growth and encapsulation process.

Initial infection starts with an infected macrophage and an intracellular bacillus that duplicates. Intracellular bacilli duplicate until lysis of first macrophage; then, approximately α extracellular bacilli are released and diffuse to nearby alveoli. Infection continues as seen in Fig 3 (exudative set) but extending to surrounding alveoli. It forms an spherical lesion with a radius that grows exponentially. This unstoppable process continues until lesion makes contact with secondary lobule membrane (*septae*), which occurs around day 51 as seen in Fig 6, where baseline values are used. This contact triggers the encapsulation process. Fibroblasts contain lesions

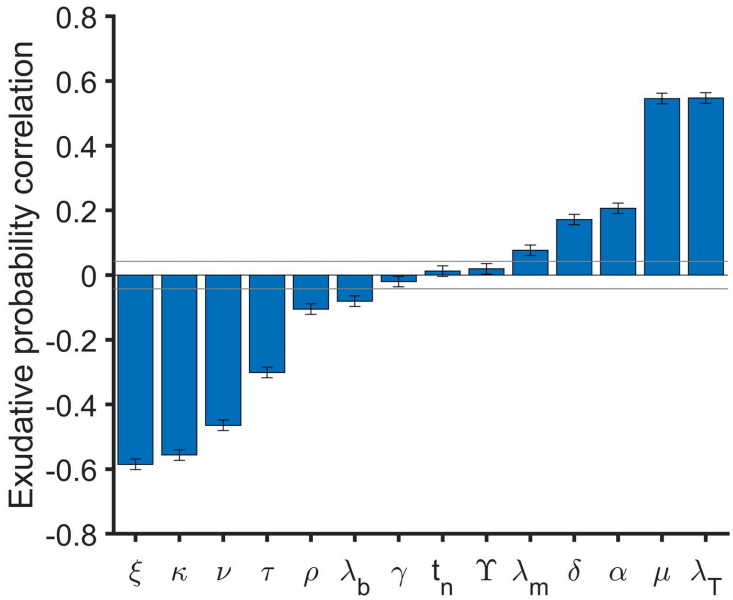

Fig 5. Partial Rank Correlation Coefficient (PRCC) between input parameters and exudative probability. PRCC between ξ, κ, ν, λ_b, γ, t_n, Y, λ_m, δ, α, τ_m and λ_T and exudative probability are shown in blue bars. Infection probability is computed at time = 200 day as the presence or not of bacilli ($b_I + b_E$). Gray horizontal lines are marked at ±0.043 to mark significance threshold for PRCC values using a significance level of 0.05. A total of 1500 parameters sets are used to explore the parameter space defined in Table 1 using Latin Hypercube Sample technique. For each set of parameters a total of 10000 simulations are done. Error bars are computed as the standard error of Spearman's correlation coefficient.

https://doi.org/10.1371/journal.pone.0239289.g005

growth and an encapsulated lesion is formed (day 73). Encapsulation process finishes around day 126 and no diffusion is observed after this time. The shape of the lesion is spherical because encapsulation process is faster than lesion growth. If encapsulation is slower (it can be seen with a different set of parameters), the shape of the lesion is oval. In fact, if fibroblasts chemotaxis is slow enough, lesions may occupy the whole secondary lobule and encapsulation is not formed.

Encapsulation time and volume of the lesions depends on fibroblast spreading. An uncertainty analysis is performed following the same procedure as in the single alveolus model. A total of 100 simulations are run, exploring 16 input parameters (Table 1). Diffusion coefficients are also varied using two parameters, D and DF. Values of diffusion coefficients are related following:

$$D_m = D_n = D_T = 10 \cdot D_b = \frac{\chi_f}{\mathrm{DF}} = \mathrm{D}. \tag{24}$$

Parameter D is varied between the double and its half of its baseline value and DF is explored between 0.5 and 4 times its baseline value. PRCC are computed between input parameters and outcome variables. Total amount of intracellular and extracellular bacilli, and macrophages are similar than the values seen in the single alveolus model, see Fig 4. In Fig 7A

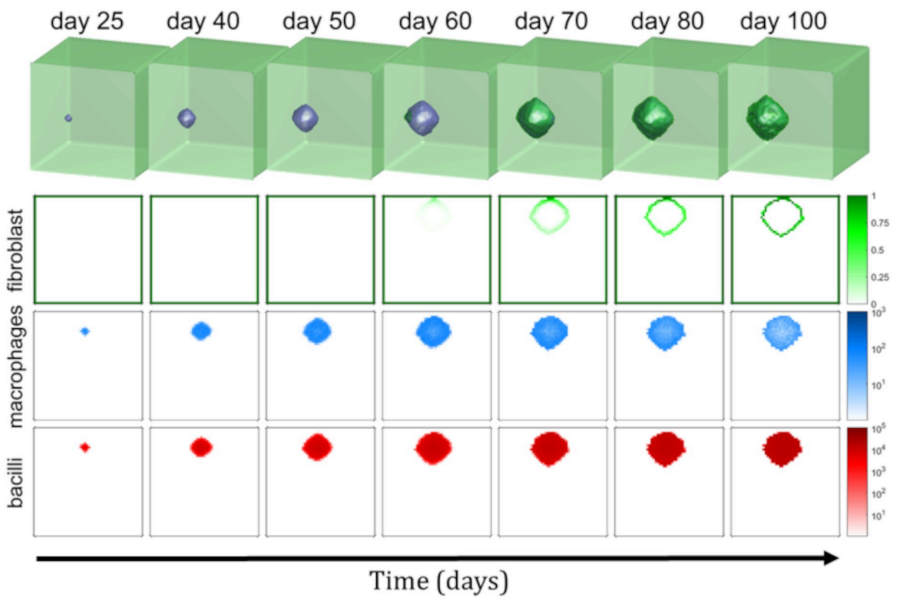

Fig 6. Evolution of 3D model. Evolution of the secondary lobule model in a cubic 50 × 50 × 50 alveoli grid. The first row is the 3D representation of the secondary lobule. Fibroblasts are shown in green and the occupied space (macrophages, T cells, dead cells, bacilli. . .) is in blue. Second, third and fourth rows represents a slice (z plane at the height of initial infection). Concentration of fibroblasts in shown in green (second row), the total number of macrophages (uninfected, infected and activated) in logarithmic scale in blue (third row) and, finally, the total number of bacilli (intracellular and extracellular) in logarithmic scale in red (fourth row). These quantities and representations can be observed at different days after initial infection, each column represents a different time of simulation.

https://doi.org/10.1371/journal.pone.0239289.g006

we observe PRCC values between volume of the lesions and the explored parameters. Lesions volume is determined by two initial parameters D, diffusion of the elements, related with lesions growth, and DF, fibroblasts diffusion, related with encapsulation velocity. If the lesions grow faster it is more difficult to encapsulate them and final lesions are bigger. In fact, the faster the encapsulation speed, the smaller the lesion. This relation is observed in Fig 7B. Size of the lesions is also determined (less influence than D and DF) by α, μ and ρ. These 3 parameters increase bacillary load, as shown in Fig 4. This causes a higher diffusion of bacilli and macrophages, thus a bigger lesion.

In Fig 7B there is PRCC between encapsulation time and input parameters. Encapsulation time is defined as the elapsed time between first time that fibroblast diffusion is non zero and the time where it becomes zero again. Parameter DF reduces encapsulation time because it increases fibroblasts diffusion and, as seen in Fig 7A, parameter D increases final lesions volume.

We have observed that bigger lesions have a larger encapsulation time, then we would expect that D increases encapsulation time. However, in Fig 7B it is seen that, contra-intuitively, it decreases. This is due to the fact that increasing D also increases fibroblasts diffusion but, in fact, as the mean value of DF is higher than 1 (it is 2.25) the increase in fibroblasts diffusion is amplified. Parameters DF and D are not the only parameters that determine lesions final volume.

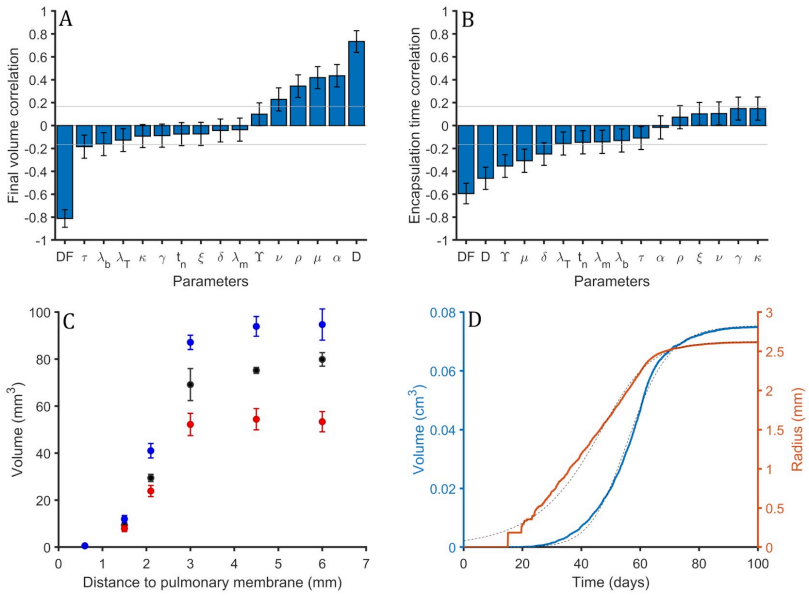

Fig 7. Sensitivity analysis of 3D model. (A) Partial Rank Correlation Coefficient (PRCC) values between volume of the final lesion and explored input parameters of the model. 100 sets of input parameters are used. Gray horizontal lines are marked at ±0.166 to mark the significance threshold for PRCC values using a significance level of 0.05. Error bars are computed as the standard error of Spearman's correlation coefficient. (B) PRCC values between encapsulation time and input parameters. 100 sets of input parameters are used. Gray horizontal lines are marked at ±0.165 to mark the significance threshold for PRCC values using a significance level of 0.05. Error bars are computed as the standard error of Spearman's correlation coefficient. (C) Distance to the nearest pleura against the volume of the resulting lesion. Each point is the mean value and standard deviation for three different simulations. Different values of D are explored. In black, baseline value, in red, baseline value divided by 2 and, in blue, baseline value multiplied by 2. (D) Volume, in blue, and radius, in orange, of the lesion using baseline values. Dotted black line is generalized logistic second degree curve approximation defined by Eqs 25–27.

https://doi.org/10.1371/journal.pone.0239289.g007

Note that Fig 7 error bars are bigger than the ones observed previously in Fig 5, due to the low number of realizations done in the secondary lobule model, because it is more computationally demanding.

Distance to the nearest pulmonary membrane strongly determines final volume of the lesion. Different simulations are carried out with different initial infection positions. As cubic secondary lobules are considered, the distance to membranes can be computed from three different axis (x, y and z). The initial infection is considered in the middle of the z axis. Then, the considered distances are 7.5 mm for each one of the sides. The distance to y membrane is fixed at 3 mm for the closest side and 12 mm for the other side. The distance to x membrane is varied for the different simulations. The considered distances are: 0.6 mm, 1.5 mm, 2.1 mm, 3.0 mm, 4.5 mm and 6.0 mm. For each different distance, three simulations are run. All simulations are repeated considering a doubled baseline value for D and using the half of its value. In Fig 7C the volume of the lesion as a function of the distance to the nearest x membrane for the three different values of D can be seen.

Two different behaviours can be distinguished. When the distance is smaller than 3 mm, a clear proportionality between distance and volume is observed. The bigger the distance the

bigger the volume. As small data is available, it can not be distinguished between exponential growth or a polynomial growth of a certain power. It is expected that volume is proportional to the cube of the distance to the nearest membrane.

When the distance is bigger than 3 mm, this behaviour changes. This is due to that the nearest distance thought x axis (that is the plotted one) is not the small one because the nearest y membrane is at 3 mm. It can be seen that the volume does not increase now with the distance. Then it can be seen a clear dependence between the volume of the lesion and the distance to the closest membrane.

In Fig 7C it is seen that the volume of the lesions increase with the diffusion (bigger values of D parameter). In blue simulations with the double of D value, in black the baseline value and in red half of the baseline value. All the colors follow the same shape. This result is also seen in Fig 7A.

Radius of the lesion can be fitted with a logistic equation. In Fig 7D lesions growth is observed (volume in blue and radius in orange) for a single secondary lobule simulation using as initial infection point an alveolus that is at 2.7 mm of the closest membrane. This growth can be approximated by a generalized logistic function [42] of second degree for the radius:

$$\frac{\mathrm{d}}{\mathrm{dt}} r(t) = v \cdot r(t) \cdot \left(1 - \left(\frac{r(t)}{r_{max}}\right)^2\right). \tag{25}$$

That can be solved imposing that there is a minimum radius (r_{min}) at a desired time (t_{min}):

$$r(t) = \frac{r_{max} r_{min}}{\sqrt{r_{min}^2 + (r_{max}^2 - r_{min}^2) \cdot exp(-2v(t - t_{min}))}}. \tag{26}$$

Catala et al [18] adjusted an agent-based model based on experiment observations of minipigs. A logistic growth was adjusted using the following parameters: r_{max} = 2.62 mm, v = 0.1324 day^{-1}, r_{min} = 0.0207 mm and t_{min} = 14 days. Imposing t_{min} = 14 days to obtain comparable results we can adjust the logistic as:

$$\begin{aligned} r_{max} &= 2.62 \text{ mm}, \\ v &= 0.1324 \text{ day}^{-1}, \\ r_{min} &= 0.0207 \text{ mm}. \end{aligned} \tag{27}$$

Comparing both models it can be seen that in this model (human) lesions are assumed to be bigger but to grow slower than minipig ones.

Discussion

Single alveolus model is useful to study macrophages-bacilli interaction observing two possible outcomes (three zones of parameters can be distinguished). In the bistability both final states and the probabilities of finishing in each one can be computed. The parameters that reduce bacillary load and exudative scenario probability are identified ($\downarrow\mu$, $\downarrow\lambda_T$, $\uparrow\xi$, $\uparrow\kappa$, $\uparrow\nu$ and $\uparrow\tau_m$) these parameters can be related with biological processes. Then, biological processes that reduce bacillary load and exudative scenario probability are a decrease of bacilli reproduction, an increase in T cells lifespan, an increase in activated macrophages lifespan, a faster activated macrophages killing rate, a faster T cells recruitment speed or a faster macrophages activation rate.

Using baseline parameters infection can not be controlled and bacilli resist inflammatory and immune response. The encapsulation process is needed (secondary lobule model) to stop

bacilli proliferation. This process controls lesion growth when immune response is not enough as seen in [7] and in the single alveolus model. The bacilli growth is not stopped but the affected alveoli are controlled not allowing to infect surrounding ones.

The modeling of such encapsulation process is closely related with the description of the growth of the granuloma. There have been several attempts to model such growing basically based in agent based models for the cells and their motion among aveoulus of the same secondary lobule [13, 17] and in the other hand reaction diffusion models for the number of cells inside the secondary lobule [43]. Our model also describes the growing of the granuloma by the use of reaction-diffusion equations, however, we incorporate in the model the saturation of such growing due to the interaction with the interlobular connective tissue by the description of the dynamics of the fibroblast.

As consequence of the approach, lesions size is affected by its position on the lung, in particular, the distance from its focus to the nearest pulmonary membrane. Other two factors that determine its size is the small diffusion velocity of the elements that is reduced by fibroblasts, collagen and other molecules present in alveoli (encapsulation velocity) and the facility that infected macrophages can diffuse to surrounding alveoli (lesion growth). Secondary lobule size and shape is important to control lesions before, therefore to control infection and not evolve to ATB. We have seen that smaller secondary lobule (smaller lungs) present smaller lesions and TB is better controlled. This could be a reason to justify the lower incidence in children than in adults observed in [44].

Volume and radius evolution of the lesions can be successfully adjusted to a generalized logistic equation as predicted by the bubble model employed in some previous empirical approaches for the growth of the granuloma [17, 18].

Further work

Next step is to build a computational lung that can reproduce TB disease that includes: bronchial tree structure [18] with a secondary lobules structure that includes pulmonary membranes and the model described before.

The details of the model can be refined including some intermediate steps that were simplified in the processes of macrophages clearance or immune activation. To study post lesion formation scenario it must be considered that when encapsulation is completed foamy macrophages can not travel outside the lesion, then process 14 (intracellular bacilli drainage) is inhibited in that lesion.

Another important point that may be taken into account is how ventilation and blood flow affect the different parts of the lung. In particular, secondary pulmonary lobules are in different conditions depending on the lung part: ventilation, blood flow, shape and size are different. These conditions may change some parameters and end in different lesions outcome.

Lesions much bigger than a single secondary lobule are obtained in active tuberculosis. In order to model the break of the pulmonary membrane, some other mechanical processes could be considered.

Supporting information

S1 Video. Encapsulation of a tuberculosis lesion in a 3D polygon secondary lobule. In green fibroblast that surrounds secondary lobule and encapsulates the lesion. In blue cells that form the lesion (macrophages, neutrophils, T cells. . .). Lesion grows till it contacts with interlobular connective tissue and it is encapsulated.
(AVI)

Author Contributions

Conceptualization: Martí Català, Clara Prats, Daniel López, Pere-Joan Cardona, Sergio Alonso.

Funding acquisition: Martí Català, Clara Prats, Daniel López, Pere-Joan Cardona, Sergio Alonso.

Investigation: Martí Català, Clara Prats, Daniel López, Pere-Joan Cardona, Sergio Alonso.

Methodology: Martí Català, Clara Prats, Daniel López, Pere-Joan Cardona, Sergio Alonso.

Software: Martí Català.

Writing – review & editing: Martí Català, Clara Prats, Daniel López, Pere-Joan Cardona, Sergio Alonso.

References

1. World Health Organization. Global Tuberculosis report. World Health Organization; 2018. Available from: https://www.who.int/tb/publications/global_report/gtbr2018_main_text_28Feb2019.pdf?ua = 1.
2. Public health emergency: WHO Global Tuberculosis Programme. TB: a global emergency, WHO report on the TB epidemic. World Health Organization; 1994. Available from: https://apps.who.int/iris/handle/10665/58749.
3. World Health Organization. The end TB strategy. World Health Organization; 2015. Available from: https://www.who.int/tb/End_TB_brochure.pdf.
4. Cardona PJ, Català M, Arch M, Arias L, Alonso S, Cardona P, et al. Can systems immunology lead tuberculosis eradication? Current Opinion in Systems Biology. 2018; 12:53—60. https://doi.org/10.1016/j.coisb.2018.10.004
5. Bermudez L, Danelishvili L, Early J. Mycobacteria and Macrophage Apoptosis: Complex Struggle for Survival. Microbe Magazine. 2006; 1(8). https://doi.org/10.1128/microbe.1.372.1
6. Cardona PJ. Patogénesis de la tuberculosis y otras micobacteriosis. Enfermedades Infecciosas y Microbiología Clínica. 2018; 36(1):38–46. https://doi.org/10.1016/j.eimc.2017.10.015
7. Gil O, Díaz I, Vilaplana C, Tapia G, Díaz J, Fort M, et al. Granuloma encapsulation is a key factor for containing tuberculosis infection in minipigs. PLoS One. 2010; 5(4):e10030. https://doi.org/10.1371/journal.pone.0010030
8. Cardona PJ, Prats C. The small breathing amplitude at the upper lobes favors the attraction of polymorphonuclear neutrophils to Mycobacterium tuberculosis lesions and helps to understand the evolution toward active disease in an individual-based model. Frontiers in microbiology. 2016; 7:354.
9. Cardona PJ. The key role of exudative lesions and their encapsulation: lessons learned from the pathology of human pulmonary tuberculosis. Frontiers in microbiology. 2015; 6:612. https://doi.org/10.3389/fmicb.2015.00612
10. Cardona PJ. What we have learned and what we have missed in tuberculosis pathophysiology for a new vaccine design: searching for the "Pink Swan". Frontiers in immunology. 2017; 8:556.
11. Verrall AJ, G Netea M, Alisjahbana B, Hill PC, van Crevel R. Early clearance of Mycobacterium tuberculosis: a new frontier in prevention. Immunology. 2014; 141(4):506–513. https://doi.org/10.1111/imm.12223
12. Haefner JW. Modeling Biological Systems. Springer; 2005.
13. Segovia-Juarez JL, Ganguli S, Kirschner D. Identifying control mechanisms of granuloma formation during M. tuberculosis infection using an agent-based model. Journal of Theoretical Biology. 2004; 231(3):357—376. https://doi.org/10.1016/j.jtbi.2004.06.031
14. Marino S, Kirschner D. A multi-compartment hybrid computational model predicts key roles for dendritic cells in tuberculosis infection. Computation. 2016; 4(4):39. https://doi.org/10.3390/computation4040039
15. Warsinske HC, DiFazio RM, Linderman JJ, Flynn JL, Kirschner DE. Identifying mechanisms driving formation of granuloma-associated fibrosis during Mycobacterium tuberculosis infection. Journal of theoretical biology. 2017; 429:1–17. https://doi.org/10.1016/j.jtbi.2017.06.017
16. Ziraldo C, Gong C, Kirschner DE, Linderman JJ. Strategic priming with multiple antigens can yield memory cell phenotypes optimized for Infection with Mycobacterium tuberculosis: a computational study. Frontiers in microbiology. 2016; 6:1477. https://doi.org/10.3389/fmicb.2015.01477

17. Prats C, Vilaplana C, Valls J, Marzo E, Cardona PJ, López D. Local Inflammation, Dissemination and Coalescence of Lesions Are Key for the Progression toward Active Tuberculosis: The Bubble Model. Frontiers in Microbiology. 2016; 7:33. https://doi.org/10.3389/fmicb.2016.00033

18. Català M, Bechini J, Tenesa M, Pérez R, Moya M, Vilaplana C, et al. Modelling the dynamics of tuberculosis lesions in a virtual lung: role of bronchial tree on endogenous reinfection. PLOS Computational Biology. 2020 May 20; 16(5):e1007772. https://doi.org/10.1371/journal.pcbi.1007772

19. Ibargüen-Mondragón E, Esteva L, Burbano-Rosero EM. Mathematical model for the growth of Mycobacterium tuberculosis in the granuloma. Mathematical Biosciences and Engineering. 2018; 15. http://dx.doi.org/10.3934/mbe.2018018.

20. Marzo E, Vilaplana C, Tapia G, Diaz J, Garcia V, Cardona PJ. Damaging role of neutrophilic infiltration in a mouse model of progressive tuberculosis. Tuberculosis. 2014; 94(1):55—64. https://doi.org/10.1016/j.tube.2013.09.004

21. Russell DG, Cardona PJ, Kim MJ, Allain S, Altare F. Foamy macrophages and the progression of the human tuberculosis granuloma. Nature immunology. 2009; 10(9):943–948. https://doi.org/10.1038/ni.1781

22. Cáceres N, Tapia G, Ojanguren I, Altare F, Gil O, Pinto S, et al. Evolution of foamy macrophages in the pulmonary granulomas of experimental tuberculosis models. Tuberculosis. 2009; 89(2):175–182. https://doi.org/10.1016/j.tube.2008.11.001

23. Cardona PJ. A dynamic reinfection hypothesis of latent tuberculosis infection. Infection. 2009; 37(2):80. https://doi.org/10.1007/s15010-008-8087-y

24. Tak T, Tesselaar K, Pillay J, Borghans JA, Koenderman L. What's your age again? Determination of human neutrophil half-lives revisited. Journal of leukocyte biology. 2013; 94(4):595–601. https://doi.org/10.1189/jlb.1112571

25. Ramos-Kichik V, Mondragón-Flores R, Mondragón-Castelán M, Gonzalez-Pozos S, Muñiz-Hernandez S, Rojas-Espinosa O, et al. Neutrophil extracellular traps are induced by Mycobacterium tuberculosis. Tuberculosis. 2009; 89(1):29–37. https://doi.org/10.1016/j.tube.2008.09.009

26. Bru A, Cardona PJ. Mathematical modeling of tuberculosis bacillary counts and cellular populations in the organs of infected mice. PLoS One. 2010; 5(9):e12985. https://doi.org/10.1371/journal.pone.0012985

27. Mahamed D, Boulle M, Ganga Y, Mc Arthur C, Skroch S, Oom L, et al. Intracellular growth of Mycobacterium tuberculosis after macrophage cell death leads to serial killing of host cells. Elife. 2017; 6: e22028. https://doi.org/10.7554/eLife.28205

28. Riley LK, Rupert J. Evaluation of patients with leukocytosis. American family physician. 2015; 92 (11):1004–1011.

29. Kramer A, Schwebke I, Kampf G. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. BMC infectious diseases. 2006; 6(1):130. https://doi.org/10.1186/1471-2334-6-130

30. Murphy J, Summer R, Wilson AA, Kotton DN, Fine A. The Prolonged Life-Span of Alveolar Macrophages. American Journal of Respiratory Cell and Molecular Biology. 2008; 38(4):380–385. https://doi.org/10.1165/rcmb.2007-0224RC

31. Gammack D, Doering C, Kirschner D. Macrophage response to Mycobacterium tuberculosis infection. Journal of mathematical biology. 2004; 48:218–42. https://doi.org/10.1007/s00285-003-0232-8

32. Krombach F, Münzing S, Allmeling AM, Gerlach JT, Behr J, Dörger M. Cell size of alveolar macrophages: an interspecies comparison. Environmental Health Perspectives. 1997; 105(suppl 5):1261–1263. https://doi.org/10.2307/3433544

33. Ting-Beall H, Needham D, Hochmuth R. Volume and osmotic properties of human neutrophils. Blood. 1993; 81(10):2774–2780. https://doi.org/10.1182/blood.V81.10.2774.2774

34. Chapman EH, Kurec AS, Davey FR. Cell volumes of normal and malignant mononuclear cells. Journal of Clinical Pathology. 1981; 34(10):1083–1090.

35. Saladin K. Human anatomy. 3rd ed. McGraw-Hill; 2011.

36. Todar K. Online textbook of bacteriology. University of Wisconsin-Madison Department of Bacteriology. 2008;.

37. Fürth R. Die brownsche bewegung bei berücksichtigung einer persistenz der bewegungsrichtung. mit anwendungen auf die bewegung lebender infusorien. Zeitschrift für Physik A Hadrons and Nuclei. 1920; 2(3):244–256.

38. Wheeler AP, Wells CM, Smith SD, Vega FM, Henderson RB, Tybulewicz VL, et al. Rac1 and Rac2 regulate macrophage morphology but are not essential for migration. Journal of Cell Science. 2006; 119 (13):2749–2757. https://doi.org/10.1242/jcs.03024

39. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of chemical physics. 2001; 115(4):1716–1733. https://doi.org/10.1063/1.1378322

40. Marino S, Hogue IB, Ray CJ, Kirschner DE. A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of theoretical biology. 2008; 254(1):178–196. https://doi.org/10.1016/j.jtbi.2008.04.011

41. McKay MD, Beckman RJ, Conover WJ. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 1979; 21(2):239–245. https://doi.org/10.1080/00401706.1979.10489755

42. Richards F. A flexible growth function for empirical use. Journal of experimental Botany. 1959; 10 (2):290–301. https://doi.org/10.1093/jxb/10.2.290

43. Hao W, Schlesinger LS, Friedman A. Modeling granulomas in response to infection in the lung. PLoS One. 2016 Mar 17; 11(3):e0148738. https://doi.org/10.1371/journal.pone.0148738

44. Seddon JA, Shingadia D. Epidemiology and disease burden of tuberculosis in children: a global perspective. Infection and Drug Resistance. 2014; 7:153.

7 Modelling the dynamics of tuberculosis lesions in a virtual lung: Role of the bronchial tree in endogenous reinfection

7.1 Summary

In this article we aim to reproduce experimental minipig data and perform some *in silico* experiments to study endogenous reinfection and the transition from LTBI to ATB. We propose a new model based on two previous models [172, 140, 36, 34].

First, the bubble model [140] which is a computational model that allows us to explain how the merging of small lesions can promote ATB in mice, and second, a mathematical model to build a computational human lung using a dichotomic algorithm [172]. The model was implemented in a bachelor's thesis and a master's thesis and, finally, it has been refined and optimized to be included in the present doctoral thesis.

In this paper, we report data of a control group in a minipig vaccine experiment [12]. Minipigs are more similar to humans than are mice [148, 70]. They are analysed to see the characteristics of TB lesions in their lungs. None of the analysed minipigs presented ATB, but they were carriers of an LTBI.

The bubble model [140] was updated according to the dynamics hypothesis [28] and adapted to the minipigs. The minipigs bronchial tree model was implemented to allow for the simulation of endogenous reinfection process. This bronchial tree was created by combining the previous human computational bronchial tree model [172, 92] with computerized tomography scan images used to reconstruct the lung surface.

The model starts with an initial group of lesions that grow until a predetermined maximum radius and that can generate new lesions. If lesions are too close, they merge, giving rise to a new lesion bigger than the old ones. The outcome of the model depends on a set of multiple parameters, most of them determined from biography and the others fitted to reproduce experimental minipig data. The initial model was built with a single initial lesion (i.e., a single initial infection site). Then, several simulations were run to explore different initial configurations with more than one lesion, but results are not conclusive.

Finally, the model is used as in silico platform to explore the transition between LTBI and ATB. To do so, ATB is defined as a set of lesions where one of them is larger than 1 cm. The most important identified parameters that determine the transition between latent regime to active regime are the inflammatory response (determines lesion size), breathing amplitude (determines how far new lesions appear) and natality index (determines how many secondary lesions are generated).

We conclude that the condition to generate an ATB is an unbalanced excessively high inflammatory response or a moderate inflammatory response combined with a close dissemination of new lesions that may be related with a smaller breathing amplitude.

7.2 Main contributions

The main contributions of this article are:

- Latent tuberculosis infection data of a minipig animal model (control group of a vaccination trial) is presented and analysed. Lesions are identified by position (x, y and z), density, distance to nearest pleura and diameter.
- A minipig virtual bronchial tree is built from a fractal model and a set of computerized tomography scan images of the lungs.
- Bubble model [140] is adapted, optimized and calibrated to fit the experimental data. The computational bronchial tree is implemented to simulate the endogenous reinfection in order to reproduce dynamics hypothesis assumptions that are submitted under falsifiability [28].
- A sensitivity analysis is performed and the most important parameters are identified with their corresponding relation with outcome variables.
- Different initial conditions are explored to identify the initial lesions among those observed in the minipig; it is not possible to determine the original lesions or their characteristics.
- Active tuberculosis disease is computationally defined. Parameters that trigger the transition from LTBI to ATB are found and related with biological and physiological processes.

7.3 Further work

Breathing amplitude, which is related with the parameter that determines new lesions spreading distance, is considered homogeneous though the bronchial tree. In humans, it is observed that this is not true; in fact, a gradient in breathing amplitude has been measured [79, 141]. A decrease in local diffusion increases the probability that lesions merge locally and ATB is generated in zones with less diffusion. Then, implementing a model with a variable breathing amplitude and checking its consequences is an interesting challenge. This is an ongoing project which has given rise to, a manuscript that is currently under preparation, with very interesting results.

In addition, the bubble model was able to reproduce ATB lesions with no bronchial tree associated in mice [140].In this article a bronchial tree was implemented and it was possible to fit LTBI lesions of a minipig model. Another interesting step would be to fit macaques ATB lesions [154] considering the dynamic hypothesis assumptions (bronchial tree dissemination). A final step would be to fit the model to human data. Nevertheless, the problem regarding human data (a part from ethical and anonymization protocols) is that computerized tomography scans are not commonly done in clinical practice and instead X-ray scans are used. In X-ray scans small lesions (< 1 cm) are not observed and only ATB cases are documented.

7.4 Article

The full article can be read online at:
Català M, Bechini J, Tenesa M, Pérez R, Moya M, Vilaplana C, Valls J, Alonso S, López D, Cardona PJ, Prats C. Modelling the dynamics of tuberculosis lesions in a virtual lung: Role of the bronchial tree in endogenous reinfection. *PLOS Computational Biology.* 2020 May 20;16(5):e1007772. `https://doi.org/10.1371/journal.pcbi.1007772`

Data set with experimental measures: `https://doi.org/10.1371/journal.pcbi.1007772.s001`

Supplementary file that includes the following supplementary sections: (1) Sensitivity analysis for transition and active set; (2) Extended sensitivity analysis; (3) Diameter-length relation in bronchial tree; (4) Details of the simplified protocol for model parametrization sensitivity; and (5) Thresholds exploration can be found at: `https:`

//doi.org/10.1371/journal.pcbi.1007772.s001

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Modelling the dynamics of tuberculosis lesions in a virtual lung: Role of the bronchial tree in endogenous reinfection

Martí Català[1,2], Jordi Bechini[3], Montserrat Tenesa[3], Ricardo Pérez[3], Mariano Moya[3], Cristina Vilaplana[4,5], Joaquim Valls[2], Sergio Alonso[2], Daniel López[2], Pere-Joan Cardona[1,4,5], Clara Prats[2]*

1 Comparative Medicine and Bioimage Centre of Catalonia (CMCiB), Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Catalonia, Spain, **2** Departament de Física, Universitat Politècnica de Catalunya, Castelldefels, Barcelona, Catalonia, Spain, **3** Servei de Radiodiagnòstic, Hospital Universitari Germans Trias i Pujol, Badalona, Catalonia, Spain, **4** Experimental Tuberculosis Unit, Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Universitat Autònoma de Barcelona, Can Ruti Campus, Edifici Mar, Badalona, Catalonia, Spain, **5** Centro de Investigación Biomédica en Red de Enfermedades Respiratorias, Madrid, Spain

* clara.prats@upc.edu

OPEN ACCESS

Citation: Català M, Bechini J, Tenesa M, Pérez R, Moya M, Vilaplana C, et al. (2020) Modelling the dynamics of tuberculosis lesions in a virtual lung: Role of the bronchial tree in endogenous reinfection. PLoS Comput Biol 16(5): e1007772. https://doi.org/10.1371/journal.pcbi.1007772

Editor: Dominik Wodarz, University of California Irvine, UNITED STATES

Received: June 5, 2019

Accepted: March 4, 2020

Published: May 20, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: CP, PJC and MC received funding from "la Caixa" Foundation (ID 100010434), under agreement LCF/PR/GN17/50300003; CV, PJC, MM, JB, MT and RP received funding from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Grup Unitat de Tuberculosi Experimental, 2017-SGR-500; CP, DL, SA, MC, JV received funding from Ministerio de Ciencia, Innovación y

Abstract

Tuberculosis (TB) is an infectious disease that still causes more than 1.5 million deaths annually. The World Health Organization estimates that around 30% of the world's population is latently infected. However, the mechanisms responsible for 10% of this reserve (i.e., of the latently infected population) developing an active disease are not fully understood, yet. The dynamic hypothesis suggests that endogenous reinfection has an important role in maintaining latent infection. In order to examine this hypothesis for falsifiability, an agent-based model of growth, merging, and proliferation of TB lesions was implemented in a computational bronchial tree, built with an iterative algorithm for the generation of bronchial bifurcations and tubes applied inside a virtual 3D pulmonary surface. The computational model was fed and parameterized with computed tomography (CT) experimental data from 5 latently infected minipigs. First, we used CT images to reconstruct the virtual pulmonary surfaces where bronchial trees are built. Then, CT data about TB lesion' size and location to each minipig were used in the parameterization process. The model's outcome provides spatial and size distributions of TB lesions that successfully reproduced experimental data, thus reinforcing the role of the bronchial tree as the spatial structure triggering endogenous reinfection. A sensitivity analysis of the model shows that the final number of lesions is strongly related with the endogenous reinfection frequency and maximum growth rate of the lesions, while their mean diameter mainly depends on the spatial spreading of new lesions and the maximum radius. Finally, the model was used as an *in silico* experimental platform to explore the transition from latent infection to active disease, identifying two main triggering factors: a high inflammatory response and the combination of a moderate inflammatory response with a small breathing amplitude.

Universidades and FEDER, with the project PGC2018-095456-B-I00; CVM has a Miguel Servet II contract funded by the Instituto Carlos III (ISCIII, CPII18/00031). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Tuberculosis is, even today, among the 10 main causes of death in the world. Despite the effectiveness of current strategies to fight the disease and those that are under development, the huge reservoir of latently infected individuals is a big hindrance in its eradication. One of the challenges inherent in this problem is that the mechanisms that cause latent infection to evolve towards active disease are not fully understood. Why will 90% of infected individuals never develop an active disease? In other words, what are the main factors that trigger an active disease in 10% of cases? We have focused our efforts on understanding the mechanisms that allow keeping infection latent, especially those related with endogenous reinfection. Since it is supposed to occur through the bronchial tree, we have designed a 3D computational model that mimics this structure, in which we have implemented an agent-based model of lesion growth and proliferation. Our results were contrasted with computed tomography measurements in latently infected minipigs, providing successful results that reinforce the essential role of endogenous reinfection through the bronchial tree in keeping infection latent.

Introduction

Tuberculosis (TB) is an infectious disease that in 2017 killed more than 1.6 million people. *Mycobacterium tuberculosis* (*Mtb*) causes TB, and this bacterium is the individual agent causing the highest mortality worldwide [1]. The World Health Organization (WHO) estimates that 25 to 30% of the population worldwide is infected with *Mtb*, and that around 10% of infected people will develop active tuberculosis (ATB) in a few years' time [2], although these percentages are being questioned and re-visited by recent studies [3]. WHO also estimates that 10 million humans developed ATB in 2017 [2].

TB infection starts at a pulmonary alveolus when *Mtb* is phagocyted by an alveolar macrophage (AM). The *Mtb* resists bactericidal mechanisms induced by AM and replicates inside the macrophage [4]. Under proper *in vitro* conditions *Mtb* replicates once a day [5]. When the intracellular bacterial load overcomes the AM's maximum tolerability, macrophage necrosis is triggered, thereby returning bacilli to the extracellular milieu. These bacilli are phagocyted by other AMs and the cycle begins again giving rise to a further increase in bacilli. The further inclusion of more AMs fails to control bacillary growth. The death of AMs triggers a local inflammatory response first, and then a specific immune response, which finally controls the infection. The end of the progressive infection leaves an encapsulated TB lesion [6]. According to the dynamic hypothesis of Cardona [7], there is a certain probability that a few bacilli will escape from the lesion, mostly inside a foamy macrophage, and start a new infection in another alveolus. This process is assumed to occur through the bronchial tree, and it is what we denote as endogenous reinfection. This includes not only new infections generated from the initial infection site, but also those originated in successive infection foci (Fig 1). An *Mtb* infection may be completely cleared by the organism [8], it may enter into a latent state in which the host is infected but not sick and cannot infect other people, called Latent Tuberculosis Infection (LTBI), or, if the immune and inflammatory responses are not well balanced, the host may develop Active Tuberculosis disease (ATB).

Systems biology and computational models are fruitful tools for increasing understanding of the processes involved in TB [1]. Recently, different models have been useful in identifying several TB key factors [9–13]. In particular, the *Bubble model* suggests that the coalescence of closed lesions is the main mechanism for the growth of lesions in animals that progress to

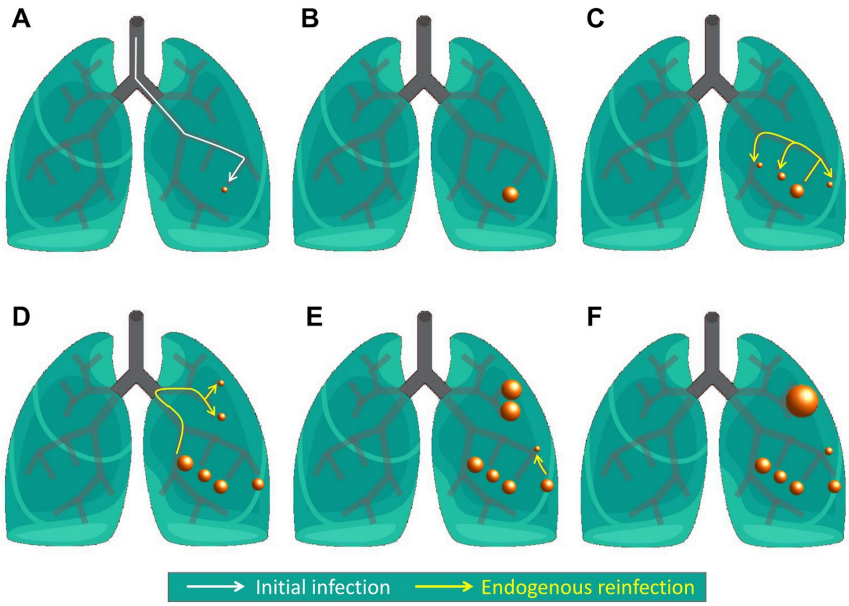

Fig 1. Main features of dynamic hypothesis. Schematic representation of initial infection (A, white arrow) lesions' growth (A–F), endogenous reinfection (C–E, yellow arrows) and coalescence of neighbouring lesions (F).

https://doi.org/10.1371/journal.pcbi.1007772.g001

ATB [14]. This model successfully explains experimental observations in mice [15]. The *Bubble model* assumes a generalised logistic (Richard's curve) [16] growth of lesions, driven by the inflammatory and immune responses, with their proliferation according to the endogenous reinfection theory, and a merging between neighbouring lesions when they are close enough. The model successfully reproduced ATB observed in C3HeB/FeJ mice, demonstrating the importance of local inflammation, lesion proliferation, and coalescence in the triggering of active disease. These results are relevant for mouse models; however, they are not easily extrapolated to humans, because of the differences between the structure of the lungs in the two species, in addition to the well-known differences in immune systems and encapsulation capacity.

Actually, the structure of the lungs may play an important role in the infection dynamics of TB. On the one hand, endogenous reinfection occurs mainly through the bronchial tree, and mice have much simpler pulmonary structure than humans, as no secondary lobular structure is found in mice (they have little or none interlobular septae) [17]. On the other hand, the encapsulation of lesions is driven by fibroblasts and fibrin from pulmonary membranes like intralobular septae. Nevertheless, mice do not possess intralobular septae and lesion encapsulation is not possible except close to the pleura [18]. In addition, the immune response of C3HeB/FeJ mice is much less effective than that of humans. In humans, a balanced Th1 immune response usually takes place after TB infection [19]. As mice's immune response is not strong and encapsulation is not possible, these animal models cannot develop an LTBI situation and all experimental observations show ATB cases [15].

Although pigs and humans share a great deal of anatomy and physiology, researchers rarely employ pigs as *in vivo* models for TB. Yet their immune system and lung structure are particularly close to the corresponding system and structure in humans. Thus, TB development in

pigs is more similar to that in humans than in mouse models [19, 20]. Minipigs are a genetically selected species, which is more convenient than other pigs for experiments in a lab, mainly for size reasons. Experimental results in TB in minipigs resemble pathological findings described in human [21–23].

In this study we aim to adapt and implement the *Bubble model* in a virtual bronchial tree in order to understand the maintenance of LTBI in minipigs. In particular, we want to test the falsifiability of the dynamic hypothesis of Cardona [7] that explains this maintenance, as well as to obtain some orders of magnitude of its dynamics. We use experimental minipig TB data to tune the model [23]. With the new model we perform several *in silico* experiments, which successfully reproduce experimental observations, and, furthermore, permit us to systematically explore the transition between ATB and LTBI.

In Materials and Methods we describe the CT experimental data, as well as the two submodels used in simulations, which correspond to the computational lung and the revised Bubble model. We finish this section providing details of the model's implementation and the methodology used for its parameterization and sensitivity analysis. Results' section starts with an analysis of the computational lung obtained. Then, it provides the results of the model's fitting to experimental data and the sensitivity analysis. Subsequent simulation series are used to explore the effect of the initial configuration and to test the transition between LTBI and ATB in minipigs. Finally, the conclusions for this study and their implications in testing the falsifiability of the dynamic hypothesis are drawn.

Materials and methods

Computer tomography measurements of LTBI in minipigs

Six female specific pathogen-free (spf) minipigs were intratracheally infected by H37Rv Pasteur strains of *Mtb* (10^3 CFU) under sedation [23]. They were euthanized twelve weeks post-infection, without having received any TB treatment. None of them had developed ATB symptoms.

Broncho-pulmonary pieces were obtained from all minipigs and analysed with multidetector computed tomography scan (CT), a GE LightSpeed VCT with high image resolution (64-slice). The morphoanatomical study was carried out with volume rendering software. For each of the minipigs, we recorded the number of lesions, their location and size, Hounsfield units, and distance to closest pleura. One of the analysed minipigs did not present any lesions; it was considered non-infected for technical reasons and excluded from the subsequent analysis.

The 5 infected animals showed 165 lesions in total, 33 ± 22 per minipig. These lesions are shown in Fig 2. The mean diameter of the lesions was 1.3 ± 0.2 mm. Lesions were located in each minipig with a mean dispersion of 16 ± 4 mm. The number of lesions and their positions were used to train the computational model [24].

Ethics statement. All ethical requirements were followed according to Directive 201/63/EU, and the protocol and procedures of the study were approved by the corresponding ethical committee on animal welfare and the Catalan Government (Permit number: 5796). All animals were euthanized at week 12 post-infection by intravenous injection of sodium pentobarbital.

Computational bronchial tree

The main novelty of this modelling approach is the use of an explicit 3D space that resembles a pulmonary bronchial tree. The design of this explicit space requires the building of a computational bronchial tree inside a certain pulmonary volume, limited by the external surface. The

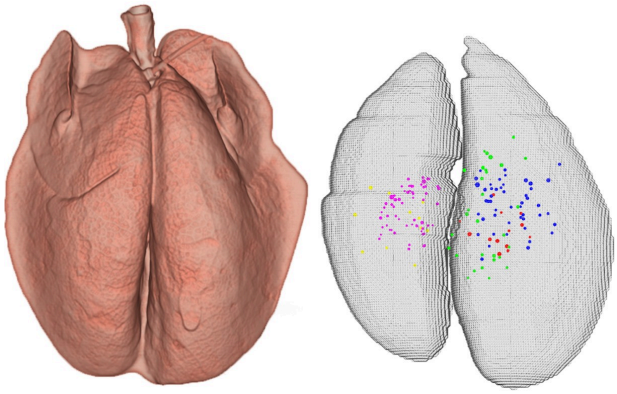

Fig 2. Summary of experimental results. Left: CT image reconstruction of a minipig's pulmonary surface. Right: 3D representation of location and size of all minipig lesions; each colour is for a different minipig (red, magenta, blue, yellow, and green).

https://doi.org/10.1371/journal.pcbi.1007772.g002

geometrical information necessary for this model can be obtained from pulmonary CT images of the studied minipigs.

Accordingly, the computational bronchial tree consists of two models: (1) an empirical model for the external lung' surface that limits pulmonary volume, and (2) an artificial iterative model of bifurcations to build a bronchial tree inside this surface. This model is deterministic, since surfaces are obtained from experimental CT measurements and the iterative model does not incorporate randomness.

Empirical model for the lung's external surface. An empirical surface model is built using CT scan data from one of the minipigs, randomly chosen. This representative surface is subsequently re-scaled according to the dimensions of others minipigs' lungs, giving rise to 5 computational surfaces that can be used to buildi the 5 different bronchial trees.

In order to build the representative pulmonary surface, we use three images of the three planes, i.e., coronal plane, sagittal plane, and axial plane. From these images, the contour line is extracted, keeping the carina (i.e., the bifurcation point of the trachea where it divides into the two main bronchi) position for purposes of reconstruction (Fig 3). The 3D reconstruction from contour lines is carried out with Matlab. All contours are normalized to 1 in order to be subsequently re-scaled with the specific dimensions of each of the 5 minipigs' lungs on the reconstruction process. Finally, the left lung is slightly rotated (5˚) so that the inter-pulmonary space is reduced.

All the lesions observed experimentally are located in the coordinate system defined by the reconstructions, taking the carina as the reference point, in order to check if they are located inside the obtained computational pulmonary surfaces. As a result, 95% of the detected lesions are inside the computed surface or in contact with pleura.

Iterative model for the bronchial tree. A bronchial tree of the conductive zone is built inside the computational pulmonary surface with an algorithm based on previous work on the human bronchial tree [25, 26]. Starting at the trachea, a set of iterative rules govern the successive bifurcations. The bronchial tree of the minipigs is assumed to be morphologically equivalent to the human one, but with smaller dimensions [27].

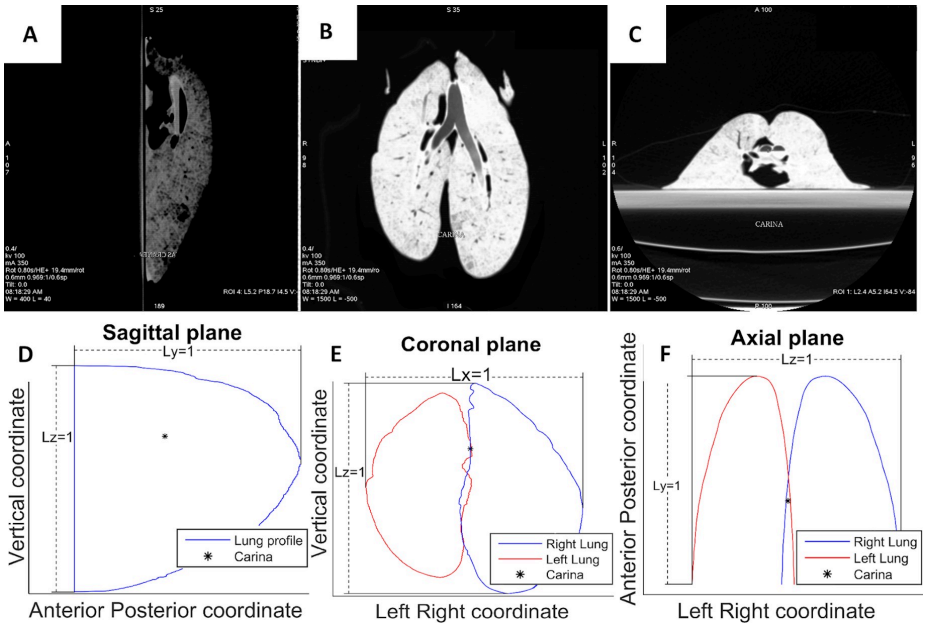

Fig 3. Normalized contour lines obtained from CT-scan images. Contour lines were obtained from CT-scan images and used for a 3D computational reconstruction of the pulmonary surface. Trachea division point (carina) is marked for purposes of reconstruction. (A) Sagittal plane CT image. (B) Coronal plane CT image. (C) Axial plane CT image. (D) Sagittal plane outline reconstruction. (E) Coronal plane outline reconstruction. (F) Axial plane outline reconstruction.

https://doi.org/10.1371/journal.pcbi.1007772.g003

Our algorithm assumes that all the divisions are bifurcations, i.e., they occur in a dichotomous way. The resulting three branches involved in a bifurcation are coplanar, and the plane that contains each bifurcation is called a bifurcation plane. The divisions are assumed to occur in successive perpendicular planes, i.e., right-left, anterior-posterior, and upper-lower. Therefore, each bifurcation plane is perpendicular to the previous plane. The first division starts at the carina and directs the new branches into the right and left lungs.

When a certain conducting airway 0 divides into conducting airways 1 and 2, the flow conservation ($Q_0 = Q_1 + Q_2$) together with Murray's law ($Q = C \cdot d^3$) [28] leads to the following relation between their diameters, d_i:

$$d_0^{\,3} = d_1^{\,3} + d_2^{\,3} \tag{1}$$

Florens et al. [29] derived a ratio of 3 for the length of a branch (l_i) and its diameter (d_i) for most of the bronchial trees:

$$l_i = 3d_i \tag{2}$$

We also studied this relation using Rozanek and Roubik's experimental data [30], obtaining a proportionality constant of 3.07 and a goodness of fit of $R^2 = 0.98$. This analysis is shown in the Supplementary material section 3 (Fig B in S1 File).

The diameters and angles of each bifurcation depend on the relative volume that each new branch supplies. We define the q_i factor as:

$$q_i = \frac{V_i}{V_0}; i = 1, 2 \tag{3}$$

where V_1 and V_2 are, respectively, the sub-volumes irrigated by conducting airways 1 and 2 after bifurcation, and V_0 is the volume supplied by the branch 0 (before bifurcation). Taking into account Murray's law, the diameters after the bifurcation are:

$$d_i = d_0 \cdot q_i^{1/3}; i = 1, 2 \tag{4}$$

Minimizing the work per unit time (associated with friction and to maintain the structure) in bifurcations, the following relations between the angles of the bifurcation and the factor q_i are obtained [31]:

$$cos\phi_i = \frac{1}{2q_i^{2/3}}\left(1 + q_i^{4/3} - (1 - q_i)^{4/3}\right); i = 1, 2 \tag{5}$$

The calculation of the ratio q_i (Eq 3) cannot be analytically evaluated. Therefore, a grid of equispaced points is created so that the number of points inside each considered volume, N_i ($i = 0, 1, 2$), is assessed and the ratio is evaluated as:

$$q_i = \frac{N_i}{N_0}; i = 1, 2 \tag{6}$$

The distance between points is initially fixed at 1 mm [25] and then reduced to 0.2 mm to increase precision and improve results.

In Fig 4 there can be seen a diagram of a bifurcation example for $q_1 = 0.6$. Using Eq 4 it can be determined that the diameter of the daughter branches are $d_1 = 0.84 \cdot d_0$ and $d_2 = 0.74 \cdot d_0$, respectively. Length is 3 times the diameter of each branch, then: $l_0 = 3 \cdot d_0$, $l_1 = 3 \cdot d_1 = 2.53 \cdot d_0 = 0.84 \cdot l_0$ and $l_1 = 3 \cdot d_1 = 2.21 \cdot d_0 = 0.74 \cdot l_0$. Bifurcation angles can be computed using Eq 5 as: $\phi_1 = 32^\circ$ and $\phi_2 = 43^\circ$.

The initial tests with these equations resulted in a bronchial tree that does not fully occupy the upper part of the pulmonary surface. Therefore, a correction is added to the angle calculation [25]. It consists of an evaluation of the mass centre of the volumes to be supplied, $\rightarrow x_{MC,V_i}$, which is projected on the bifurcation plane. We will denote with ψ_i the angle between the original branch and the projected mass centre. Experimentally, bifurcations with an angle greater than $\pi/2$ were not observed [32]. Therefore, the corrected angle φ_i is:

$$\varphi_i = min\left\{\frac{\phi_i + \psi_i}{2}, \frac{\pi}{2}\right\}; i = 1, 2 \tag{7}$$

Branches with a diameter lower than 0.5 mm are considered terminals as areas those branches that escape from the pulmonary surface. The model for the building of successive bifurcations is summarized in Table 1.

The *Bubble model*: an update

The *Bubble model* is an agent-based model in which the agents are the lesions. The dynamics of the lesions are driven by three processes: a generalised logistic growth, the reinfection process that permits the generation of new infection focuses, and the coalescence between neighbouring lesions.

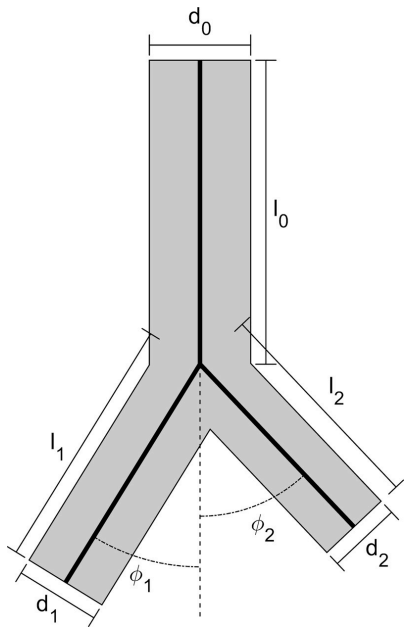

Fig 4. Bifurcation diagram. Bifurcation of 0 branch into two (1, 2) daughter branches. The cabal ratio for branch 1 is: $q_1 = 0.6$. Length is 3 times the diameter of each branch as may be seen in Eq 2. Diameter relations are obtained from Eq 4, as $d_1 = 0.84 \cdot d_0$ and $d_2 = 0.74 \cdot d_0$. Angular values are computed using Eq 5, as $\phi_1 = 32^\circ$ and $\phi_2 = 43^\circ$.

https://doi.org/10.1371/journal.pcbi.1007772.g004

The *Bubble model* was originally designed and calibrated to describe the dynamics of tuberculous lesions in mice with an active disease [14]. The spatial structure was not relevant for mice, due to the dimensions and the relatively simple structure of mouse lungs, and taking into consideration the size of lesions for the active disease. This model is updated and implemented as follows.

Lesion growth. We model a lesion as a sphere whose spatial position (3D coordinates, in mm), radius (in mm), and age (in days) are variables. Lesions are firstly detected when their radius is $r_{min} = 0.075$ mm (smaller lesions cannot be identified). This occurs after approximately $t_{min} = 14$ days from the initial infection. Then when a lesion is created it remains "silent" for 14 days before it is initialized with a r_{min} radius. The model employs a generalised logistic growth of the radius of the lesions as follows:

$$\frac{dr_i(t)}{dt} = \nu_i \cdot r_i(t) \cdot \left[1 - \left(\frac{r_i(t)}{r_{max}}\right)^2\right] \tag{8}$$

where r_i is the radius of the lesion, ν_i is the parameter that sets the maximum growth rate, and r_{max} is the maximum radius. The parameter ν_i is modelled as a Gaussian variable with mean value ν and standard deviation $\nu/3$. Therefore, each lesion grows at a slightly different velocity at each time step. From experimental data it is known that around the 28th day a 2 mm lesion reaches its limit [22]; therefore, ν is estimated as $\nu = 0.3$ day^{-1} ($\sigma_\nu = 0.1$ day^{-1}).

Table 1. Summary of the model for building the computational bronchial tree of each minipig.

The normalized pulmonary surface is specifically re-scaled for each minipig, taking into account the measured dimensions.
A bronchial tree is built inside each computational pulmonary surface.
The bronchial tree starts at the end of the trachea, taking the trachea diameter and carina location of each minipig as reference.
The bronchial tree is a tubular structure, and non-terminal branches split in a dichotomous way.
The three branches implied in a division are coplanar (bifurcation plane).
In each division, the pulmonary territory is divided in two subregions by the plane that is perpendicular to the bifurcation plane, following the direction of the mother branch.
The bifurcation plane of the first division is the vertical one that separates the right and the left lungs.
The bifurcation planes are perpendicular from one generation to the following.
Once the bifurcation plane is defined, the ratio $q_i = V_i/V_0$ is numerically evaluated by a grid of equiespaced points with a precision of 0.2 mm, so that $q_i = N_i/N_0$ (N_i and N_0 are the number of points contained in each subvolume).
The parameters of each bifurcation are evaluated using Eqs 2, 4, 5 and 7.
Branches that escape the pulmonary surface and those with a diameter of less than 0.5 mm are considered terminals.

https://doi.org/10.1371/journal.pcbi.1007772.t001

Lesion proliferation. The multiplication of the number of lesions is caused by endogenous reinfection. In this way, a *mother* lesion generates new *daughter* lesions from day 14 to day 28. The original reinfection probability function [14] includes two terms: (1) a linearly increasing term with the radius of the mother, and (2) a linearly decreasing probability with *mother* lesion age [22]. The second term is slightly modified in order to allow the generation of new *daughter* lesions from *mother* lesions older than 28 days, with a small non-zero probability:

$$P(t)dt = \rho \cdot \frac{r_i(t)}{r_{min}} \cdot e^{-\alpha(a_i-14)^n} dt, a_i \geq 14 days \tag{9}$$

where ρ, α, and n are parameters that define the probability profile, a_i is the age of the lesion, in days, and r_{min} is the minimum radius at which lesions are identified. Fig 5 shows original [14] and modified (Eq 9) models with $\alpha = 0.035$ day^{-n} and $n = 1.63$ with $r_{max} = 1$ mm and $\rho = 0.10$ day^{-1}. The values of α and n are fixed to ensure that the area under the two curves are equal and to minimize the difference between the experimental results and the linear model. The reinfection probability initially grows exponentially with the lesion radius; however, for longer times the curve decays exponentially because of encapsulation and calcification.

In the original model for mice, the location of new lesions is selected from a probability which decreases with the distance. In the current version for minipigs the distance is modelled in the same way, but considering the bronchial distance between two terminals instead of the geometric distance between two points. The model does not explicitly assume that the dissemination occurs exclusively through the aerial part, but it could include a possible recirculation through the adjacent circulatory or lymphatic systems as well. Then, possible locations are determined by the bronchial tree terminal positions:

$$P(i \rightarrow j) = \frac{e^{-\beta d_{ij}}}{\sum_j e^{-\beta d_{ij}}} \tag{10}$$

where $P(i \rightarrow j)$ is the probability that a lesion appears at a terminal j due to a *mother* lesion at terminal i, and β is the dispersion parameter that determines the spreading. Fig 6A shows the mean distance of the appearance of new lesions as a function of the dispersion parameter. Fig

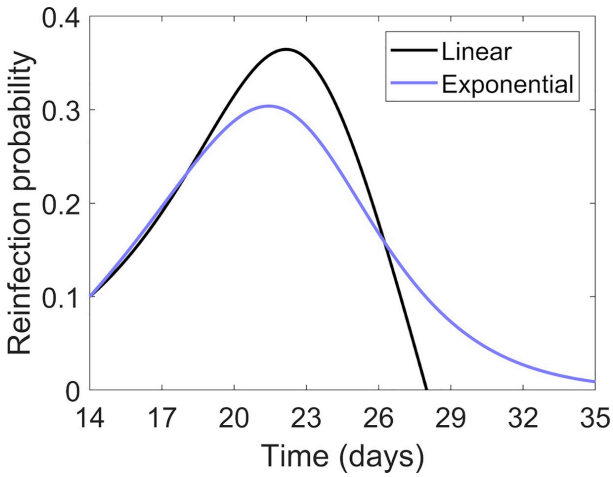

Fig 5. Reinfection probability. Reinfection probability for a lesion with r_{max} = 1 mm and ρ = 0.10 day^{-1}. In black, original *Bubble model* [14]; in blue, the updated model considering an exponential decrease during the control phase. The area under the two curves is equivalent.

https://doi.org/10.1371/journal.pcbi.1007772.g005

6B shows the distribution of distance from daughter to mother lesions for β = 0.08 mm^{-1}, which is slightly different from the theoretical distribution (Eq 10) due to spatial quantization.

Lesion coalescence. Coalescence occurs when the distance between two neighbouring lesions is shorter than the largest radius. Unlike the original model [14], the properties of the resulting lesion are weighted according to the coalescing lesions. The new radius is:

$$r_{new}{}^3 = r_i^3 + r_j^3 \tag{11}$$

where r_i and r_j are the original radii of the two lesions. We employ the radius to calculate the

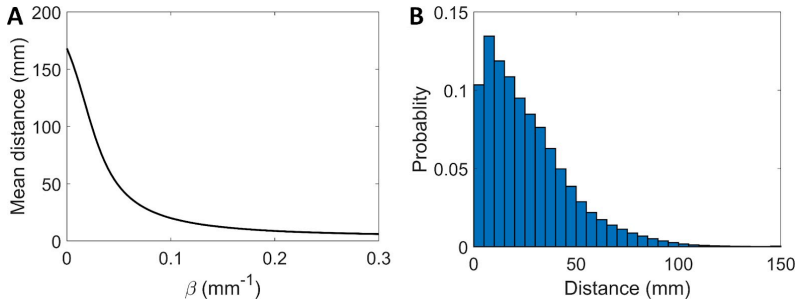

Fig 6. Lesion spreading distance. (A) Mean distance where lesions appear as a function of the dispersion parameter, $\beta \in [0, 0.3]$ mm^{-1}. Due to the quantization of the space, for $\beta > 0.3$ mm^{-1} the probability that the lesion appears at the nearest terminal is very high and the value is a constant. (B) Density of probability of new lesions spreading with β = 0.08 mm^{-1}.

https://doi.org/10.1371/journal.pcbi.1007772.g006

weights:

$$\omega_k = \frac{r_k^3}{r_{new}^3}; k = i, j \tag{12}$$

which are used to calculate the new age and position of the lesions

$$a_{new} = \omega_i \cdot a_i + \omega_j \cdot a_j \tag{13}$$

$$\vec{x}_{new} = \omega_i \cdot \vec{x}_i + \omega_j \cdot \vec{x}_j \tag{14}$$

Coalescence process is considered every 10 time steps to reduce computational cost and to ensure that the results are not modified. There are no significant differences when one time step is used.

Implementation and fitting

The computational models are implemented in Matlab. First, the bronchial trees of the 5 analysed minipigs are generated to obtain 5 virtual lungs. Next, we obtain the terminal locations and the distance between terminals through the bronchial tree. Finally, the evolution time of the lesions are calculated using the updated *Bubble model* implemented in the virtual lungs.

The model depends on a set of input parameters. It is not a deterministic model so the same set of parameters can have different outcomes. But there is a strong dependence between input parameters and outcome variables. Parameters $t_{\min}$, $r_{\min}$, α, n, and v are adjusted as detailed above, according to the available information. The simulated total time, $T_{\max}$, is equal to the experimental time (i.e., 84 days). The number of time steps is fixed as 1000 in order to ensure model stability. There are three remaining parameters to be fitted:

- **Mean maximum radius, $r_{\max}$** is the mean radius achieved by non-merged lesions following a generalised logistic growth. It is measured in mm. To account for biology variability, each lesion has a different maximum radius. We use a Gaussian distribution of standard deviation $\sigma_{r\max} = 0.2 \cdot r_{\max}$ around the mean value.
- **Natality index, ρ** is proportional to the probability of triggering an endogenous reinfection process. It is measured in day^{-1}.
- **Spreading parameter, β** determines spreading of the lesions through the bronchial tree due to endogenous reinfection process. It is measured in mm^{-1}.

The complete model reproduces the evolution of an initial distribution of TB lesions in a 3D computational lung. Each single simulation consists of 5 independent runs, one for each virtual lung. The outcome results of the model are listed in Table 2.

Simulation scheduling. In order to assess the precision of the results obtained from the model, the variability of the different outcomes (Table 2) is measured. 25000 simulations are run to determine this variability. As mentioned above, in each simulation the model is run for the 5 virtual lungs that correspond to the 5 minipigs. A total of 125000 runs are done. This analysis is useful to learn whether the different results obtained with differing initial parameter simulations are significant or not.

The variability (error, E) of a given outcome, OC, is defined as:

$$E(OC, N) = \left| \frac{M(OC) - m(OC, N)}{M(OC)} \right| \tag{15}$$

where $m(OC,N)$ is the mean value of a given outcome after performing N simulations and M

Table 2. Outcome variables of the model.

Outcome variable	definition
Number of lesions	Number of lesions with a diameter greater than 0.9 mm (resolution threshold). These are called observable lesions.
Mean diameter	Mean diameter of the observable lesions. This is measured in mm.
Dispersion	Mean distance between observable lesions and their coordinates centre. This is measured in mm.
Coalescences	Number of merging processes that occurred during simulation.
Disease indicator	Logical index that indicates the presence or not of lesions larger than 10 mm.
Diameters histogram	Histogram of the diameter of the observable lesions. Edges of the histogram can be controlled.
Coordinates histogram	Histogram of the positions of the observable lesions. Edges of the histogram can be controlled.
Disease coordinates	Coordinates of the lesions larger than 10 mm. These are measured in mm taking the carina as origin.
Diameters dispersion	Standard deviation of the diameter of observable lesions. This is measured in mm.
Wounded volume	Total volume occupied by observable lesions. This is measured in cm^3.
Time till disease	Elapsed simulation time till a larger lesion of 10 mm is found. This is measured in days.

https://doi.org/10.1371/journal.pcbi.1007772.t002

(OC) is the limit of $m(OC,N)$ when N is large enough, in this case $N = 25000$. The outcome values follow a Gaussian distribution, sometimes truncated due to quantization effects. Running simulations with different input parameters, $S = \{\rho, \beta, r_{\max}\}$ and different number of simulations (N), a dependence between this variability (E), the targeted outcome (OC), the number of simulations (N), and the used set of parameters (S) is seen. This error is inversely proportional to the square root of the number of simulations. Then, Eq 15 can be rewritten as:

$$E(OC,N,S) = \left|\frac{M(OC,S) - m(OC,N,S)}{M(OC,S)}\right| = \frac{A(OC,S)}{\sqrt{N}} \tag{16}$$

where $A(OC,S)$ is the proportionality constant that depends on the outcome variable (OC) and the given set of input parameters (S). Table 3 presents the values of $A(OC,S)$ for different sets of parameters and outcomes.

From this simulation schedule, we conclude that 500 simulations for each dataset ensures an error (variability) lower than 4% in most of the outcome values and different sets of

Table 3. Values of proportionality constant, $A(OC,S)$, for different outcomes and sets of input parameters.

		Latent set	Transition set	Active set
Input parameters	ρ (day^{-1})	0.120	0.045	0.045
	β (mm^{-1})	0.08	0.10	0.15
	r_{max} (mm)	0.68	6.50	10.00
Proportionality constant	**Number of lesions**	0.62	0.70	0.38
	Mean diameter	0.07	0.32	0.20
	Dispersion	0.10	0.88	3.38
	Coalescences	4.70	0.58	0.61
	Disease percentage	−−−	0.58	−−−

Values are approximated using $N = 25000$ simulations.

https://doi.org/10.1371/journal.pcbi.1007772.t003

parameters. Only in a few cases where the mean numeric value of the outcome is very small do we observe errors larger than 4%, but never larger than 20%.

Simplified protocol for model parameterization. We designed a protocol for parameterization, to find the set of parameters, $S = \{\rho, \beta, r_{\max}\}$, that best fit experimental data. We start our simulations with a single lesion at the mass centre of the experimentally observed distribution. To fit the three remaining parameters, we built three error functions to evaluate the agreement of three outcomes of the computer simulations with experimental data:

1. Error in number of lesions (*NLE*): relation of the number of lesions observed experimentally, $N(E)$, and the final number of lesions obtained from a given simulation, $N(S)$:

$$NLE = \frac{|N(E) - N(S)|}{N(E)} \tag{17}$$

2. Error in distribution of diameters of the lesions (*DE*), relation of the diameter histogram, $HD_i(E)$, of the experimentally observed lesions with the corresponding simulation outcome ($HD_i(S)$), n_{bins} being the number of bins in the histograms:

$$DE = \frac{1}{N(E)} \sum_{i=1}^{n_{bins}} \left| HD_i(E) - HD_i(S) \frac{N(E)}{N(S)} \right| \tag{18}$$

3. Error in spatial location of lesions (*SE*), comparison of the experimental ($HS_{i,j}(E)$) and the numerical ($HS_{i,j}(S)$ histograms of the spatial coordinates):

$$SE = \frac{1}{3 \cdot N(E)} \sum_{j=x,y,z} \sum_{i=1}^{n_{bins,j}} \left| HS_{i,j}(E) - HS_{i,j}(S) \frac{N(E)}{N(S)} \right| \tag{19}$$

These error functions will subsequently be used as objective functions to be minimized in the parameterization process.

The dependence between the objective functions and the input parameters is shown in Table 4. In this table, green means that the objective function is sensitive to this input parameter and that we can observe its minimum; orange means that it is sensitive but no minimum is observed; red means no sensitivity. Detailed plots are shown in supplementary material, section 4 (Fig C and Fig D, both in S1 File).

Given these results, we design the following process to fit the three parameters: (1) β is fitted by minimizing *SE*; (2) then, $r_{\max}$ is fitted by minimizing *DE*; (3) finally, ρ is fitted by

Table 4. Sensitivity of the error functions (NLE, DE, and SE) to the three parameters explored (β, ρ and r_{max}).

Error function	β	ρ	r_{max}
Number of lesions error (NLE)	IM	IM	IM
Diameter error (DE)	NM	NA	IM
Spatial error (SE)	IM	NA	NA

In green (IM), error functions that are affected by these parameters for which a minimum can be identified; in orange (NM), those that are affected but do not present a distinguishable minimum; in red (NA), those functions that are not affected by the parameter.

https://doi.org/10.1371/journal.pcbi.1007772.t004

minimizing *NLE*. This process is repeated iteratively to finally fit the three parameters jointly, because a change in one parameter can slightly move the minimum position of the three error functions.

Sensitivity analysis. The sensitivity analysis is performed using as initial infection a single lesion on the mass centre of the experimental lesions observed experimentally in each minipig. The sensitivity analysis is performed for 5 variables: β, r_{max}, ρ, ν, and T_{max}. A set of 11 simulations with different parameter values is designed: one simulation with all the parameters in their default values, five simulations increasing one-at-a-time parameter by 10%, and five simulations decreasing one-at-a-time parameter by 10%. We compute an ANOVA test analysis for 4 outcomes by comparing the results obtained with the original set of parameters and those obtained with each new parameter combination [33]. An extended sensitivity analysis is performed using [34] methodology; this sensitivity analysis is shown in Supplementary material section 2 (Fig A in S1 File).

Results

Computing the bronchial tree model: Properties and fitting with experimental data

The surface of the minipig lungs is obtained from CT images. The 5 bronchial trees are generated inside the surfaces; see an example of computer generated bronchial tree in Fig 7. The largest tube is the trachea and at each division the alveolar diameter and length are reduced, according to the algorithm. The branches ramify and occupy all the pulmonary territory.

Fig 8 gathers an analysis of one of the computed bronchial trees, which is representative of what is observed in other cases. Although we have no experimental information about the exact geometry of minipig bronchial trees, we can analyse the main characteristics of virtual lungs generated and discuss their reliability with existing knowledge and data in general terms. We find a Gaussian distribution around 45˚ for the bifurcation angular distribution, which is similar to that observed experimentally in humans (Fig 8A). The mean bifurcation angle is 47.7˚. The imposition of no angles higher than 90˚ gives rise to a peak at 90˚. As expected, the terminal branch generation is greater than the non-terminal. Most of the branches end at generation 14. The highest generation observed in our computed virtual bronchial trees is between 29 and 31, depending on the dimensions of the lung, which are slightly different for each minipig. The mean terminal diameter is 0.39 mm with a minimum value of 0.1 mm and a maximum of

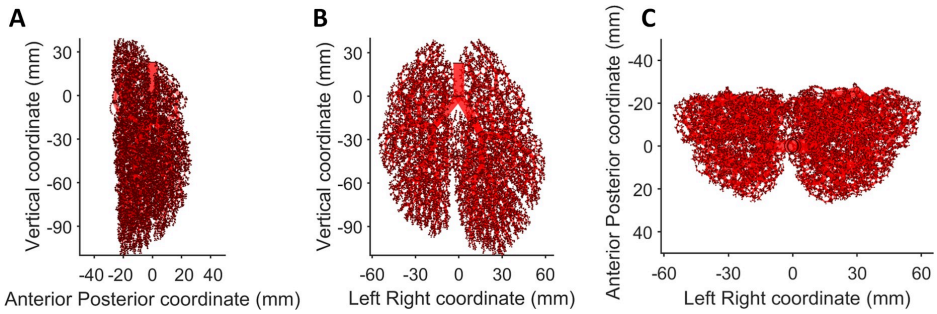

Fig 7. Computational minipig bronchial tree. It is represented as a tubular structure. Each conducting airway is represented by an empty cylinder. Three axial planes are shown: (A) Sagittal plane, (B) Coronal plane, and (C) Axial plane.

https://doi.org/10.1371/journal.pcbi.1007772.g007

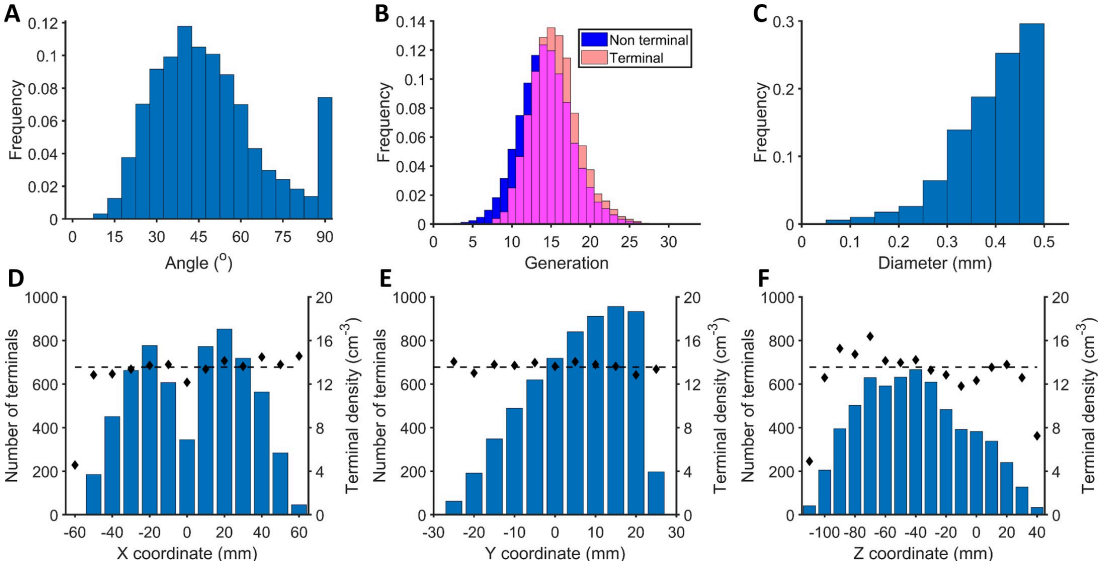

Fig 8. Computational bronchial tree analysis. (A) Angular distribution of bronchial tree bifurcations. The angle is measured between the mother branch and the new one. (B) Generation distribution; in red, the terminal branches, in blue the non-terminal, and in pink the intersection between them. (C) Diameter distribution of the terminal branches. (D) Terminal distribution (blue) and density (black dots) along X coordinate. (E) Terminal distribution (blue) and density (black dots) along Y coordinate. (F) Terminal distribution (blue) and density (black dots) along Z coordinate.

https://doi.org/10.1371/journal.pcbi.1007772.g008

0.5 mm. The minimum terminal diameter observed is determined by net spacing. Fig 8 also shows the location of terminals and their density along the bronchial tree. This particular mini-pig's bronchial tree has 6267 terminals and its volume is 462 cm^3; this means a density of 13.6 terminals cm^{-3}. As shown by black points in Fig 8D–8F, terminal density is isotropic, and thus constant along the 3 spatial axes. This suggests that the bronchial tree is reasonable, because it would be able to supply oxygen to all terminals equally distributed around the pulmonary territory. Terminal distributions show how the pulmonary volume is distributed along the 3 axes (i.e., the lower the pulmonary volume in a certain zone, the fewer the absolute number of terminals).

Most terminals (90%) have the closest terminal between 1.0 mm and 3.2 mm. However, in a few zones with low pulmonary space assigned we may observe great variations in the mean value and each terminal has another terminal at a distance of between 0.1 mm and 6.1 mm.

Fitting the model with simplest initial assumptions

The updated *Bubble model* is used to simulate the evolution of an initial infection in a computational lung, taking into account experimental results. Experimental CT images showed the final state of the infection; we do not have direct information on its initial location. Coordinates and diameters datasets are analysed to determine whether they follow a Gaussian distribution. These Gaussian distributions would be indicative of a group of initial lesions that were generated from a single infection process and that would have evolved from this common origin with the same mean growth rate. Nevertheless, after a one-sample Kolmogorov-Smirnov test we rejected the null hypothesis, which is that the data were part of a standard normal distribution.

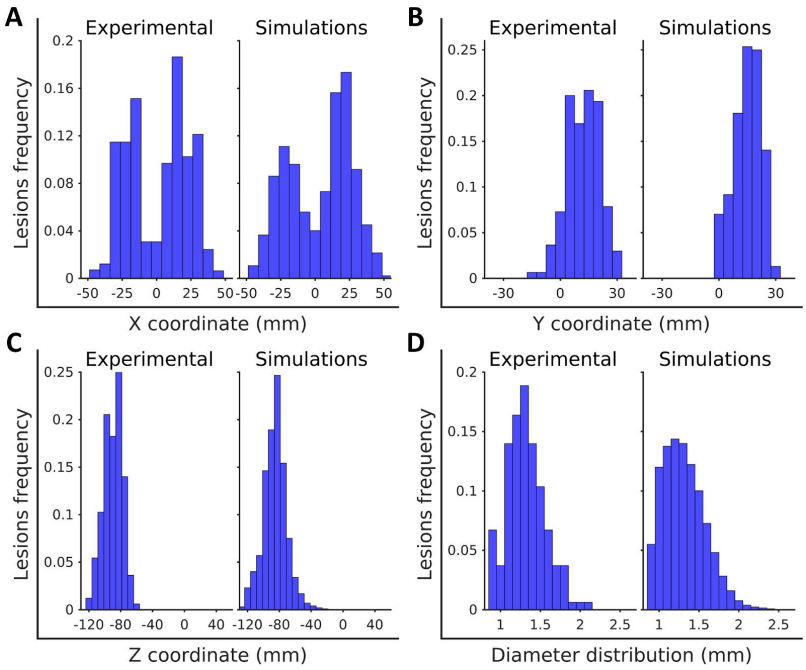

Fig 9. Comparison of experimental and computationally obtained distributions of lesion location and size. Computational distributions were obtained considering as initial infection one lesion in the mass centre of the observed lesions. The set of parameters used is: $\rho = 0.13$ day^{-1}, $\beta = 0.08$ mm^{-1}, and $r_{max} = 0.68$ mm. (A) Coordinate X (Left—Right) histogram comparison. (B) Coordinate Y (Anterior–Posterior) histogram comparison. (C) Coordinate Z (Vertical) histogram comparison. (D) Diameter distribution histogram comparison.

https://doi.org/10.1371/journal.pcbi.1007772.g009

The lack of reliable information about the initial infection entails the need for a blind assumption. Following the law of parsimony, we assume the simplest initial configuration, and alternative possibilities will be explored later on. We use as initial infection for the simulations a single lesion located at the nearest terminal to the mass centre of all the lesions, given the final distribution shown by CT images. The set of parameters is adjusted to minimize the errors as explained in Materials and Methods. followed by the performance of 500 independent simulations of each minipig's configuration (i.e., a total of 2500 simulations). The set of parameters obtained after the minimization of the objective functions (Eqs 17–19) is $\rho = 0.13$ day^{-1}, $\beta = 0.08$ mm^{-1}, and $r_{max} = 0.68$ mm, which corresponds to $NLE = 0.089$, $DE = 0.283$, and $SE = 0.299$.

In Fig 9 we show the comparison between the experimental and the simulated outcome distributions of spatial coordinates X, Y, and Z, and of lesion diameters. A two-sample Kolmogorov-Smirnov test of the four pairs of distributions show that there are not significative differences in any of the cases with a signification level of 0.05. Therefore, the distributions resulting from the numerical simulation successfully reproduce the experimental observations.

Simulations show that coalescence of lesions is nearly non-existent, on average less than one coalescence per minipig. This result is in agreement with Prats et al. [9] and Marzo et al. [15], who presented coalescence as a mechanism essential to the evolution towards an active

Table 5. Sensitivity analysis for the set of parameters: S = {ρ, β, r_{max}} = {0.12 day^{-1}, 0.08 mm^{-1}, 0.68 mm}.

		Input parameters				
		β	r_{max}	ρ	ν	T_{max}
Outcome variables	Number of lesions	0.137	0.092	<0.001**	<0.001**	<0.001**
	Mean diameter	0.003**	<0.001**	0.475	0.106	<0.001**
	Dispersion	<0.001**	0.613	0.089	<0.001**	<0.001**
	Coalescences	0.024*	0.004**	<0.001**	<0.001**	0.754

Each value is the minimum of the p-values from ANOVA test comparing 500 runs of the original set of parameters and the set of parameters where one parameter is increased or decreased by 10%. Numbers marked with

* present statistically significant differences with $p<0.05$ and numbers marked with

** with $p<0.01$.

https://doi.org/10.1371/journal.pcbi.1007772.t005

disease from a latent infection. A lack of coalescences, therefore, would be a control indicator of the latent infection.

Furthermore, simulation results show that the process of endogenous reinfection is crucial to understanding how lesions appear at different locations in the lungs. The use of a computational bronchial tree for driving such reinfection produces spatial distributions which resemble the experimental cases. This process has been shown to be a key factor in maintaining a latent infection inside big mammals like minipigs.

Sensitivity analysis

Table 5 shows the results of the sensitivity analysis, with the minimum value for the ANOVA test between the increased and the decreased parameters. This analysis reveals that the number of lesions is strongly related with ρ, ν, and T_{max}. It was not obvious that the number of lesions would be related with the growth velocity; however, when lesions grow faster there is an increase in the likelihood that the process of endogenous reinfection will generate new lesions.

The mean diameter varies with parameters β, r_{max}, and T_{max}. The inflammatory response is the cause of lesion growth, so relations with r_{max} and T_{max} are expected. The results also show that the dispersion parameter, β, slightly affects the mean diameter. A smaller dispersion parameter causes lesions to be closer, thereby increasing the chance of a coalescence event. In fact, as seen in extended sensitivity analysis for the explored parameter space, r_{max} and β are the two parameters that affect the mean diameter value most. An increase in one of these parameters increases mean diameter value.

According to this analysis, T_{max} is the only parameter that is not related with the resulting number of coalescences. All other parameters affect the coalescence processes; nevertheless, a counter-intuitive result is that the dispersion parameter is not the most strongly related. One may expect the dispersion parameter to be the parameter that would affect the coalescence process most because it is the one that determines coalescence spreading, and then determines the distance at which new lesions appear.

The sensitivity analysis evidently depends on the initial set of parameters. Nevertheless, we have employed other sets of parameters to carry out sensitivity analyses, providing equivalent results. These analyses are shown in the Supplementary material section 1 (Table A and Table B).

Analysing the effect of initial conditions

After confirming that the model with the simplest assumption for the initial conditions is good enough to explain the experimental results, we explored the possibility of improving the

Table 6. Initial configurations explored with the model using experimental data.

Initial infection	Parameters set			Errors			
	ρ (day^{-1})	β (mm^{-1})	r_{max} (mm)	*NLE*	DE	*SE*	$SE/SE_{control}$
Mass centre (control)	0.134	0.071	0.67	0.015	0.25	0.30	1.00
Coordinate centre	0.134	0.070	0.66	0.026	0.23	0.34	1.14
Biggest lesion	0.129	0	0.68	0.006	0.18	0.74	2.45
Two biggest lesions	0.102	0	0.68	0.015	0.18	0.73	2.42
30% biggest lesions	0.084	0	0.69	0.009	0.17	0.77	2.58
Densest lesion	0.123	0	0.69	0.003	0.19	0.73	2.43
Two densest lesions	0.110	0	0.68	0.026	0.17	0.73	2.44
30% densest lesions	0.084	0	0.68	0.004	0.16	0.78	2.58
Density>150HU	0.078	0	0.68	0.020	0.17	0.79	2.62
One random lesion	0.227	0.170	0.54	0.027	0.63	0.25	**0.82**
Two random lesions	0.122	0.129	0.58	0.008	0.35	0.23	**0.77**
One random terminal	0.212	0.160	0.55	0.011	0.59	0.71	2.36
Two random terminals	0.129	0.150	0.59	0.016	0.40	0.71	2.37

The first column shows the criteria for choosing which of the measured lesions were assumed as initial infection. The following columns show the parameter values that minimized errors (ρ, β, r_{max}) and the values of the three errors obtained (*NLE*, *DE*, and *SE*). The last column compares the spatial error objective function (*SE*) with the one from the control simulation. HU: Houndsfield units.

https://doi.org/10.1371/journal.pcbi.1007772.t006

agreement between the model and experimental measurements by testing different initial distributions of lesions. Table 6 shows the 12 initial configurations analysed, in addition to the previous one. We explore the choice of one or more lesions from CT data as the initial infection using location, size, and density criteria, as well as different random choices. For each initial distribution, the set of parameters is adjusted to minimize the objective functions as explained in Materials and Methods. Table 6 shows the parameter values that minimize errors as well as the corresponding values for each of the explored initial distributions.

The results of this analysis, shown in Table 6, do not provide a conclusive criterion for distinguishing those lesions that belonged to the initial infection. Nevertheless, they corroborate that the final lesion distribution is strongly related with the initial infection distribution, since the objective function that is most affected is that of spatial error (*SE*). The distributions that assume as initial infection one or two random experimental lesions provide better spatial error objective function results; however, they give rise to larger values for *DE*.

In some cases, the value of the spatial parameter that minimizes the spatial error function is $\beta = 0\ mm^{-1}$. This value permits a macrophage to travel to all other terminals not taking into account the distance through the bronchial tree. In consequence, the final distribution becomes one that follows all terminal spatial distributions and is not related with the particular initial distribution. Therefore, in these cases the initial distribution is related neither with density nor diameter.

From latent to active tuberculosis: *In silico* experiments

Mathematically we define a case of active disease as one with numerical simulations providing a lesion larger than 1 cm in diameter [35]. The model is designed to reproduce experimental results from latent tuberculosis in minipigs. Therefore, no trigger of disease is observed in any of simulations with the fitted parameters. The following set of simulations is designed to explore the parameter space, looking for those zones leading to active disease.

The parameter space is delimited by $\beta \in [0, 0.2]$ mm^{-1}, $r_{max} \in [1, 5]$ mm, and $\rho \in [0.02, 0.2]$ day^{-1}. We used equidistant points, 11 for β, 10 for r_{max}, and 4 for ρ. We explored a total of 440 points. We ran 2500 simulations for each point of this parameters space, and 500 for each minipig virtual lung. The initial infection configuration is set as the control (i.e., one lesion in the mass centre of the measured lesions' distribution). Finally, we define an *Active Disease Index* as the frequency of active cases among the total number of *in silico* experiments for each point of the parameter space.

We show in Fig 10 the obtained *Active Disease Index* for each of the explored points. This index increases with any of the 3 parameters, r_{max}, β, and ρ. Three different zones of the parameters space can be defined: a latent zone, where most of the simulations result in a latent infection (green colour in Fig 10), an active disease zone, where most of the simulations derive into an active disease (red colour in Fig 10), and a transition zone, where both dynamics are possible (from light green to dark orange in Fig 10).

The parameter r_{max} is related with the effective inflammatory response in a broad sense; the greater the effect of the inflammatory response, the bigger the lesions and the greater the likelihood of developing an active disease. This result is in agreement with experimental observations [15] and with other biology systems approaches [14]. Of course, the effective dynamics of inflammatory response can be modulated by local properties such as oxygen concentration or macrophages' availability, among others, which are not explicitly considered by the model.

Parameter β is the dispersion parameter; a higher value of β corresponds to lower dispersion inside the lung, which can be a consequence of a lower breathing amplitude. Therefore, a

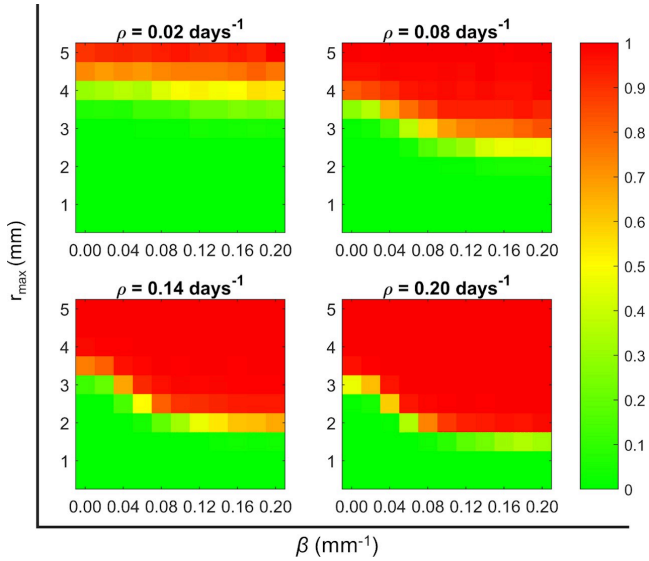

Fig 10. Active Disease Index for different sets of parameters. Exploration of parameters space (r_{max}, β and ρ) to see the fraction of *in silico* experiments that present an active TB disease. The colour is proportional to this frequency; green colour means most of the cases remained latent, red colour means that most of the cases derived into an active disease, and intermediate colours mean that both dynamics are possible.

https://doi.org/10.1371/journal.pcbi.1007772.g010

low breathing amplitude appears again as a possible cause for the appearance of big lesions, as was previously described in the literature [9]. It has to be taken into account that breathing amplitude changes from one lobe to another one, i.e., it is wider in lower lobes. Therefore, the conditions that facilitate the development of an ATB would vary from one lobe to another. Our model shows that active disease can be triggered by a high inflammatory response or due to a moderate inflammatory response combined with a small breathing amplitude. Nevertheless, at this point the virtual lungs are considered to be homogeneous, i.e., we are not considering variations of parameters along the lungs' structure.

After this exploration, the sensitivity analysis is extended to check if sensitivity of the model depends on the set of parameters used. In addition to the default parameters that belong to the latent infection zone, we chose two combinations of parameters: one representative of the active disease zone and the other of the transition zone; see Fig 10. A latent TB parameter set shows that the space parameter, β, determines mostly the lesions dispersion; however, a parameter set of the transition zone (r_{max} = 6.5 mm, β = 0.1 mm^{-1}, ρ = 0.05 day^{-1}) shows that the space parameter affects mostly the number of lesions, mean diameter, and coalescences, and does not affect the dispersion of the lesions. This can be seen in Supplementary material, section 1.

Discussion

Limitations and further work

In this modelling approach we have followed the law of parsimony (Occam's razor) [36], trying to find a simple solution for complex problems such as TB infection dynamics in lungs. The level of complexity was chosen according to the questions to be addressed. This method was developed specifically to submit the main assumptions of the dynamic hypothesis to falsifiability testing. Therefore, the current model includes the most important steps of TB infection evolution suggested by this hypothesis: endogenous reinfection, lesion growth, and coalescence. These processes are supposed to capture the essence of TB dynamics in lungs, but of course they are not the only ones [6]. In fact, no model could be complete [37]. This also allows the use of a fewer number of parameters, when compared with other systems biology approaches to the same problem [1], and thus provides more robustness to the fitting.

The principal novelty of this model is the implementation of the bubble model in an explicit space like the bronchial tree in order to simulate the endogenous reinfection processes. Nevertheless, it still has a few limitations that should be mentioned, the most important being the following:

- Exogenous reinfection is not yet considered in this model. Its incorporation may change the outcome when simulating an ATB infection, as it acts as a new mechanism to generate new infection focuses. Nevertheless, the experimental data used in this study were obtained under conditions that prevented exogenous reinfection. Therefore, the inclusion of this mechanism should be supported by experimental designs that allow it.

- The bronchial tree model is absolutely deterministic, for now. In the future we expect to add some random noise in this algorithm in order to obtain different bronchial trees from a single pulmonary surface. This will be useful for analysing the role of specific bronchial tree properties in TB evolution as well as to account for heterogeneity sources.

- Infection spreading parameters are uniform in each virtual lung. Nevertheless, breathing amplitude is not constant, but varies from lower and middle lobes (wider amplitude) to upper ones (lower amplitude). Breathing amplitude is probably related with lesion spreading; then, higher values of β would better fit the local behaviour of less dispersion in the upper lobe, while lower values of β would be appropriate for describing the spreading in the lower and

8 Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance

8.1 Summary

In this article, two compartmental models are presented in order to study the coexistence between human species and *Mtb*. The main goals are to study the role that the higher female resistance and tolerance to TB played and how the substitution from "ancient" lineages to "modern" ones was carried out.

Two sets of parameters are used to simulate the population behaviours in Paleolithic (longer lifespan but fewer births per woman) and Neolithic (more births per woman but shorter lifespan). Two lineages are compared, the "modern" lineage that appeared around 43,000 years ago that is more contagious and virulent and the "ancient" one that was there from 70,000 years ago [50, 20].

Firstly, a compartmental SEIR model to study the tolerance and resistance factor in female was built. As gender is distinguished, there are 8 possible states (TB-Susceptible, Exposed, Infectious or Recovered $\times$ 2 genders). Possible transitions and their rates were obtained from bibliography. As 63% more cases are observed in male (62% of the cases are male and 38% of the cases females), we adjusted a factor for females to simulate higher tolerance (slower progression and less probability to pass to infectious state) and more resistance (less chance of dying if infected). The rest of the equations are the same for both genders.

If tolerance and resistance factors are one (no difference between genders) we observe that population and TB are extinguished using paleolithic parameters and there can not be coexistence because TB is too lethal and transmissible. Then, these factors are crucial to understand TB and humans coexistence. However, using this factor as "modern" lineage is more contagious and severe population in paleolithic is also extinguished . Then coexistence in paleolithic is likely due to the female protection and "ancient" lineage. In neolithic, coexistence is possible with and without female protection and although "modern" lineage is extra-contagiousness, population growth is heavily affected.

A second model is used to study the substitution of the "ancient" lineage by the "modern" lineage. It is a SEIR-based model where both infections are possible. Then, a total of 16 states are possible (4 "modern" lineage SEIR states × 4 "ancient" lineage SEIR states). Independently of initial conditions, "modern" lineage surpasses "ancient" as it is more contagious.

A discrete resolution of the two models is used to estimate clearance and extinction factors and final incidences. These include a certain degree of randomness, in contraposition to the deterministic essence of the continuous solving of the models. Here, the number of individuals in each state is a non-negative integer. Transitions between states are random Poisson number with Poisson parameter the transition rate

8.2 Main contributions

The main contributions of this article are:

- A SEIR model is built to study coexistence between TB and humans. The difference between genders is implemented to simulate the higher resistance and tolerance that females have to tuberculosis in comparison to males.

- This female effect is crucial in the tuberculosis coexistence with humans in the Paleolithic. If this female factor did not exist, coexistence would not have been possible.

- Different simulations with different circumstances (different lineage and different social conditions) are performed to determine what the evolution of the population would have been the evolution of the population. It is seen that "modern" lineage surgence in the Paleolithic was not possible because it would have extinguished the communities were it appeared.

- We have built a model to see the coexistence of the two lineages. In the Neolithic the "modern" lineage prevailed over the "ancient" one despite initial conditions. Around double the number of "modern" cases compared to "ancient" ones were observed in the equilibrium state. There was also a small fraction of individuals that present both diseases but it was more than 10 times smaller than any other lineage.

- A discrete version for each model was built. We used them to compute extinction and clearance rates and to observe the conditions that determine the final states. As expected from continuous models, the "modern" lineage was not able to survive in the Paleolithic; it would cause the extinction of the population or else disappear from the population.

8.3 Further work

From this model one important factor to be studied is the effect of smear-positive individuals and their percentage. It has been observed that sick individuals that present sputum smear negative (SN) have a 15 times lower mortality compared to those that present sputum smear positive (SP) [146]. It is not well known what the proportion of individuals that had SP was, or how this proportion would have affected the population.

Same authors (Pere-Joan Cardona, **Martí Català**, and Clara Prats) recently submitted a new paper: "The origin and maintenance of tuberculosis is explained by its subclinical course, the Neolithic revolution being the trigger for its devastating deadly drift" in *Frontiers in Microbiology* that is currently under review. The main goal of this study was to evaluate the minimum rate of protection needed to induce mild SN TB in order to maintain the consensual annual population growth established in the Paleolithic and Neolithic societies, and also to determine the impact of "ancient" and "modern" *Mtb* lineages

8.4 Article

The full article can be read online at:
Cardona PJ, Català M, Prats C. Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance. *Scientific reports*. 2020 Jan 8;10(1):1-20. `https://doi.org/10.1038/s41598-019-56769-1`

Supplementary information may be found at: `https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-019-56769-1/MediaObjects/41598_2019_56769_MOESM1_ESM.pdf`

SCIENTIFIC REPORTS
nature research

OPEN

Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance

Pere-Joan Cardona[1*], Martí Català[2] & Clara Prats[3]

Current data estimate the origin of *Mycobacterium tuberculosis* complex (MtbC) infection around 73,000 years before the common era (BCE), and its evolution to "modern" lineages around 46,000 BCE. Being MtbC a major killer of humanity, the question is how both species could persist. To answer this question, we have developed two new epidemiological models (SEIR type), adapted to sex dimorphism and comparing coinfection and superinfection for different MtbC lineages. We have attributed a higher resistance/tolerance to females to explain the lower incidence noted in this sex, a better health status in the Paleolithic compared to the Neolithic, and a higher dissemination of "modern" lineages compared to "ancient" ones. Our findings show the extraordinary impact caused by "modern" lineages, provoking the extinction of the groups infected. This could only be overcomed by an unprecedented population increase (x20 times in 100 years) and helped with the protection generated by previous infection with "ancient" lineages. Our findings also suggest a key role of female resistance against MtbC. This data obliges us to rethink the growth population parameters in the Paleolithic, which is crucial to understanding the survival of both MtbC and humans, and to decipher the nature of human female resistance against TB.

Tuberculosis (TB) is a major threat to humankind. Indeed, it has been estimated that this disease has caused 1,000,000,000 deaths in the last 200 years[1]. Despite the major efforts made to control it, including the world emergency declared by the WHO in 1993[2], TB still is a challenge, causing 1.5 million deaths last year alone. In addition, the incidence of this disease is declining only slowly despite the considerable global efforts invested in trying to improve its prevention, diagnosis and treatment[3]. Several groups have attempted to determine the origin of *Mycobacterium tuberculosis* Complex (MtbC) in order to better understand its highly virulent nature. In contrast to the classic theory whereby it originated as a zoonotic infection, evolving from an ancestor of *Mycobacterium bovis*[4], current data support the opposite. Thus, MtbC's most recent common ancestor (MRCA) emerges around 73,000 years before the common era (BCE), originating specifically in the so-called anatomically modern humans (*H. sapiens*) from an environmental mycobacterium[5–7]. This is linked to the onset of controlled fire use, and thus smoke exposure and increased physical contact among individuals around the fire[8]. Interestingly, this MRCA differs from "ancient" and "modern" lineages in a context of low population densities around 70,000 and 46,000 BCE, respectively[6,7]. During this period, known as the Middle Paleolithic, the most populated continent (Africa) had sustained a human presence since roughly 2,000,000 BCE in a context of several glaciations and their very balanced lifestyle[9]. Organized into small tribes of around 50 individuals, these humans were nomadic hunter gatherers, had a good health status thanks to a lifestyle based on a varied diet, low work intensity and moderate exercise. This resulted in a life expectancy of around 33 years[10,11]. However, this population did not grow due to low birth rates and the difficulty in surviving up to the age of 15 (only 57%) as the cost of children was high[11]. Once they had reached the age of 15, 67% of these humans lived to an age of 45 or older[12]. This explains why, even in the context of a good lifestyle, the human population remained stable for hundreds of thousands of years, with

[1]Unitat de Tuberculosi Experimental, Institut de Recerca Germans Trias i Pujol (IGTP), Universitat Autònoma de Barcelona, CIBERES, Badalona, Catalonia, Spain. [2]Comparative Medicine and Bioimage Centre of Catalonia (CMCiB). Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Catalonia, Spain. [3]Escola Superior d'Agricultura de Barcelona, Departament de Física, Universitat Politècnica de Catalunya (UPC)-BarcelonaTech, Castelldefels, Catalonia, Spain. *email: pj.cardona@gmail.com

a population growth of around 0.003%/year or less[13,14]. This scenario is radically different to the one observed during the increased growth population explosion linked to the Neolithic revolution, of around 0.1%/year[13], when human activity became more farming-based, which led to a more sedentary life style and higher birth rates, but also a lower quality of life and increased mortality[10,15].

In this context, the question remains as to how a devastating disease like TB did not eradicate humankind given the low population density of the Middle Paleolithic. We have also addressed this question by considering the particularities of the bifurcation that occurred 46,000 BCE with the onset of "modern" lineages after loss of the Tbd1 gene region[5]. Various studies have tried to discern the biological differences between these lineages and modern Mtb strains, using mainly *M. africanum* (lineages 5–7) and lineage 1 for the "ancient" strains and lineages 2–4 for the "modern" ones. Portevin *et al.*[16] compared the innate response triggered in macrophages and dendritic cells by different strains from both lineages and concluded that "ancient" ones induce a higher pro-inflammatory profile when infecting macrophages and dendritic cells. These results have been interpreted as a virulence/immune response trade-off for "ancient" strains as bacilli are discovered early by the immune system and have less chance to progress, including extrapulmonary dissemination, which contrasts with the situation found for modern strains[17]. However, the current interpretation does not include this trade-off as this pro-inflammatory profile can, in fact, enhance the infiltration of neutrophils into Mtb-infected lesions, thus resulting in a better chance to quickly progress to active TB[18]. Bold *et al.*[19] discovered a relevant characteristic of *M. africanum*, namely the larger size of the bacilli. This fact could indeed be a trade-off between progression to disease versus dissemination as it might hamper the production of small aerosols, which are best able to effectively reach the alveolar macrophages[20], thus limiting the spread of these lineages.

Thus, the trade-off for "ancient" MtbC lineages is the higher probability of progression to active TB at a cost of being less able to disseminate through the population explaining why "ancient" lineages can only be found in certain geographical locations[21]. The fact that the appearance of "modern" lineages (46,000 BCE) coincided with a significant increase in the population (from 10^4 to 10^6 individuals) in Asia, while remaining constant in Africa (around 10^6 people)[22,23], indicates that some sort of population explosion occurred in Asia before the Neolithic, as proposed by several authors[24–26], giving support to its dissemination.

The current interpretation of the coevolution between humankind and MtbC is based on the hypothesis that the mechanism of infection of MtbC was originally based mainly on the induction of latent infection, with a late progression towards active disease (i.e. prolonged latency) of more than a generation, with younger and more susceptible individuals subsequently becoming infected[27]. In accordance with these criteria, Zheng *et al.*[28] adapted the treatment-free model of TB transmission[29,30] to a population of 100 individuals. In their model, these authors did not consider the special risk of recovered TB cases subsequently going on to develop active disease again, which is at least seven-times higher than in latently infected individuals[31], although they did consider the exogenous reinfection process. They concluded that maintaining the persistence of MtbC required a high progression to disease of up to 50%, which far exceeds the value of 5–10% considered nowadays[32], justifying this by assuming a progressive increase in resistance acquired by humanity over time. The achievement of a considerable level of resistance in a community requires a large proportion of the population to be submitted to a mortality high enough to select innately resistant subjects, as shown by several authors[14,33,34]. The question is, could the Paleolithic population afford such a high progression rate?

From an eco-immunological viewpoint, host resistance reduces the parasite concentration (i.e. exploitation) below the optimum for the bacilli, thus leading to a counter-adaptation to enhance exploitation[35]. This could lead to a negative scenario for humans as parasites are better able to adapt due to the larger size of their populations and shorter generation times, which would lead to an "arms race co-evolution"[36]. This would lead to a theoretically indefinite escalation in resistance against exploitation, which would most probably defeat the host. In this regard, Bergelson[37] proposed a sort of cyclical dynamic, with escalation followed by a reduction in competition[35]. Equally, there is another mechanism, namely tolerance, that can benefit both organisms. Tolerance is the ability to maintain fitness when hosting a high parasite concentration. Moreover, tolerance increases the fitness of the host, tends to be fixed, and at the same time allows persistence of the infection[38]. Overall, it appears that natural selection tends to favor combinations of resistance and tolerance (high tolerance and low resistance, or low tolerance and high resistance, or intermediate values of both), a correlation that has been demonstrated in plants[39].

In light of these factors, we have developed a new SEIR model based on previous versions[28,29] that includes known factors such as exogenous reinfection and the increased susceptibility to progression to TB in individuals who have already suffered TB disease and have recovered[31]. We have also included the constant bacillary expulsion, together with the decrease in immunity observed with time[40–42], avoiding the concept of "once infected always infected and protected" proposed by Stead[43], which is currently no longer valid[42,44]. In this model, we have considered the most recent data available concerning fast progression and reactivation determined for infection with "modern" lineages[45,46], together with data on natural cure and mortality in untreated HIV-negative patients[47]. We have also taken into account the health conditions of humans in the Paleolithic and Neolithic, which might have affected changes in protection mechanisms, understood to be a combination of tolerance and resistance mechanisms. In addition, we have considered this mechanism to explain the different incidence in the two sexes (globally, 64% of cases are found in males and 36% in females)[3]. Indeed, the lower TB incidence in women is a matter of controversy as it has been attributed largely to cultural and socioeconomical inequalities against women[48]. However, this concept is currently being challenged, with the opposite, i.e. an inequity of health services among men, being claimed[49]. The idea of some type of natural protection in women due to biological mechanisms was first raised several years ago[50].

Application of these concepts may help us to understand how "modern" MtbC lineages, with their current dissemination capacity, were able to persist until the present without eliminating humanity. In this regard, we have also established a co-infection model in order to study the replacement of "ancient" lineages with their "modern" counterparts found nowadays.

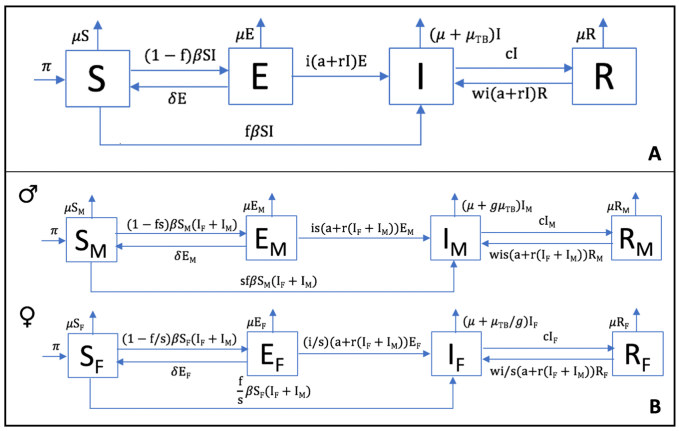

Figure 1. SEIR model. Each compartment refers to the set of individuals by disease. status: Susceptible, Exposed, Infected, Recovered. New-born individuals are assumed. susceptible. A TB infection can remain latent (**E**), or can directly develop into infectious. active TB (**I**). The latent TB infection can become active through endogenous reactivation or exogenous reinfection. Patients with active TB can naturally recover (**R**) becoming non-infectious. Latent infected persons (**I**) can drain the bacilli, lose the immunity and become susceptible (**S**). Recovered persons can relapse to active TB through endogenous reactivation or exogenous reinfection. Picture **A** shows the generic model, and picture **B** the one that considers the sexual dimorphism (TBOREX), where new-born entrance π depends on the fertility of females.

To the best of our knowledge, this is the first time that the concepts of tolerance/resistance and sex differences have been taken into consideration when trying to understand TB epidemics, in addition to co-infection with different lineages. Our findings should be taken into consideration in the current analysis of global epidemics, in which 1/3 of TB cases are not even identified. This is a paramount factor that should be addressed to stop the pandemic and to improve the survival of all subjects at high risk of developing TB.

Results

The irruption of "modern" lineages caused a dramatic impact in the Paleolithic. We have designed a compartmental mathematical model (TBOREX) (TB, Origen and Sex) based on five differential equations to describe the dynamics of the evolution of MtbC infection in the population, based on previous models[28–30,51] (Fig. 1 and Table 1). The standard scenario modelled is a human group of 50 persons in which a single infectious male is initially included. Figure 2 shows the progression of both MtbC lineages in the conditions of the Paleolithic and Neolithic periods.

In the case of a Paleolithic community with the "ancient" strain, the logarithmic scale shows two periods. The first of these, known as the "attack" period, occurs during the first 10 years after one infectious case enters the communities, thus representing initial dissemination. This phase is characterized by a sudden increase in the number of exposed and infectious cases, which is linked to a decrease in the susceptible population (Fig. 2A). The global population remains stable at a cost of a high annual mortality (Fig. 2I), reaching levels of 1000 deaths per 100,000 inhabitants, together with an initial reduction in incidence and prevalence (Fig. 1E). A second wave then appears, with this wave following a growth that stabilizes the epidemic 80 years post-challenge (Fig. 2E) with a mortality slightly higher than 1000 deaths/100,000 inhabitants and an incidence and prevalence of 4508 and 9305 per 100.000 inhabitants, respectively (Figs. 2E and 3). Interestingly, there is a constant growth in the recovered population, with a "stable" period being reached at 100 years post-challenge. The susceptible compartment undergoes a negative progression that stops with a decrease in incidence, subsequently increasing slightly until the infectious compartment increase again, finally stabilizing at 5.1 cases (Fig. 4).

Infection with a "modern" lineage, with a much higher dissemination power, rapidly extinguishes the susceptible compartment under Paleolithic conditions (in three years), as shown in Fig. 2B. Infectious cases reach a plateau, which is followed by a parabola-like kinetic that peaks at more than 10 cases (Fig. 2B) as a result of the rapid progression of infected and recovered cases. This results in a decline in the population, after stabilization of the epidemic, at an extremely high incidence of 18,329 cases/100,000 inhabitants and a mortality of 4819 (Fig. 2F,J, and 3). Note that, under Neolithic conditions, the total population grows extremely rapidly during infection with "ancient" lineages, up to 351 persons, thus multiplying sevenfold in 100 years (Fig. 2C). In contrast, for a "modern" infection, the population roughly doubles in the same period (Fig. 2D).

Higher population growth in the Neolithic allows the persistence of "modern" lineages. When analysing the Neolithic scenario, it appears that the first phase occurs as in the Paleolithic (Fig. 2), except for the higher number of susceptible individuals as a consequence of the higher population growth. In the following

Parameter	Values		Sources
	Paleolithic	Neolithic	
Annual population growth rate	1%	2.6%	13,26,52
Number of births per fertile woman	2.8	4.7	
Number of births per fertile woman/year (λ)	0.078	0.128	
Natural mortality/year (μ)	1/33	1/26.5	73,74
Mortality/year caused by TB (μ_{TB})	0.12	0.15	47
Infected people per case/year (e)	10(A)/20(M)		67
Fast progression (f)	0.099(A)/0.0825(M)	0.1238(A)/0.1031(M)	45
Reactivation from infection (a)	f · 0.3		
Bacillary drainage and immunity reduction (δ)	0.1-a-r		40,41
Reduced progression due to immunity (i)	0.1		69
TB natural cure (c)	0.33		47
Increased progression in Recovered (w)	7		31
Male/Female TB tolerance (s) (g)	55/45		50

Table 1. Parameters and references.

phase, between 10 and 100 years, growth is exponential in both cases, especially during infection with the "ancient" lineage. However, this is a non-realistic scenario as, after 1000 years, the population reaches more than a billion if the birth rate is not limited by a logistical growth with a specific carrying capacity. In addition, given the higher dissemination capacity and lower clearance of "modern" lineages, it can be assumed that these lineages displaced "ancient" ones and markedly restrained the potential growth of the Neolithic population. This will be analysed below, with the co-infection model. In this case, the incidence of infectious persons increases to a significant 10% of the population (Fig. 2H), with a slightly increasing trend.

"Modern" lineages are better able to persist in the population. The continuous resolution of the model provides a unique global view of the progression which, unfortunately, does not fully reflect reality as we are working with very low numbers. In addition, the continuous resolution of the model's equations allows the average dynamics of the system to be observed but does not account for the inherent variability between different communities due to their limited size. In particular, a number of people between 0 and 1 in the infectious compartment is halfway between completely different situations that correspond to the absence (0) or presence (1) of individual people in this compartment. This is why we decided to work with the TBOREX discrete resolution, where values are transformed into natural numbers using random numbers, as detailed above (Supplementary Fig. 2). Figure 5 shows the percentage of runs in which the TB-affected compartments (E, I and R) disappear after running the program 10,000 times. Indeed, this simulation series emulates the behaviour of epidemics in 10,000 independent communities using the same model and parameters, but with a certain degree of randomness to account for inherent variability. The results reflect how "modern" lineages are cleared from communities less often (almost half) than the "ancient" ones in periods of 100 years, thus confirming that a higher capacity for dissemination is a key factor for persistence of these epidemics. The distribution of the final TB incidence found for the Paleolithic scenario, before extinction of the community due to a "modern" lineage infection, is of special interest given the wide range of possible final incidences.

Fast progression, immunity, latency and higher susceptibility in recovered subjects are required to maintain epidemics. Figures 6 and 7, and Supplementary Figs 3–6, analyse the impact of neutralizing different factors that are intrinsic to the nature of MtbC infection, such as fast progression (f) and reactivation from latent infection (a), since these factors have been used to explain the persistence of MtbC[27,28]. The results show the extreme importance of fast progression, without which epidemics disappear in all cases due to an almost 100-fold reduction in incidence (Supplementary Figs 3–6). Neutralization of endogenous reactivation (a) seems to have a lower impact in terms of incidence, prevalence and mortality (around twofold), but, curiously, when looking at the clearance factor, it appears to have a similar impact (increase of about 3.5-fold) as the neutralization of fast progression (Fig. 4). This illustrates that both factors are essential for the maintenance of TB epidemics.

The second most important factor is the higher susceptibility in recovered cases (w), neutralization of which reduces the incidence by three- and 10-fold in "ancient" and "modern" lineages, respectively (Supplementary Figs 3–6). This factor has a lower impact on the clearance of epidemics (less than twofold) (Fig. 4).

The sensitivity analysis confirmed this view (Fig. 8 and Table 2). In this case we have also added immunity (i), an increase in which (and thus decrease in the protection conferred by immunity) is an important factor for maintaining the incidence, as is the increase in natural mortality (μ). An increase in both these factors is also related to the increase in clearance (Fig. 8B). In contrast, the birth rate (λ) works in the opposite way by decreasing both incidence and clearance.

The importance of these factors changes with time, as can be seen from Supplementary Figure 7. Focusing on the TB-related classes (E, I and R), in the attack phase (i.e., the first 10 years), fast growth (f) and number of contacts (e) have particular relevance, subsequently becoming irrelevant once the epidemic has stabilized. Reactivation (a) also shows a similar correlation pattern in the particular case of infectious subjects. Birth rate

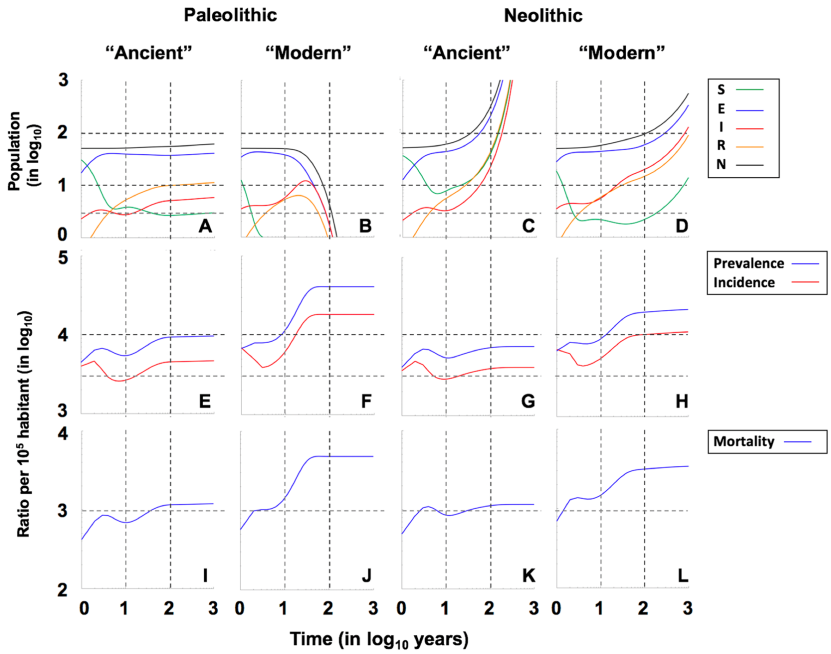

Figure 2. Evolution of the population in the continuous model TBOREX. Pictures show the projections for 1000 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$). Evolution is drawn under Paleolithic (**A,B,E,F,I,J**) and Neolithic (**C,D,G,H,K,L**) life conditions, in relation to the MtbC lineage infection, "Ancient" (**A,E,I,C,G,K**) or "Modern" (**B,F,J,D,H,L**). Continuous lines in pictures **A** to **D** show the total population (**N**) (**black**), Non-infected susceptible (**S**) (**green**), Infected (**E**) (**blue**), TB patients Infectious (**I**) (**red**) and previous TB patients Recovered (**R**) (**orange**). Figure C is truncated as the levels grow exponentially to reach final outcomes of $>10^9$ persons. Continuous lines in pictures **E** to **H** show the Incidence (**red**) and Prevalence (**blue**) of TB cases (**I**) per 100.000 inhabitants. Continuous lines in pictures **I** to **L** show the Mortality caused by TB cases per 100.000 inhabitants. Dotted lines represent reference values. We have considered 2 and 10 persons (population), 3.000 and 10.000 (incidence and prevalence), 300 and 1.000 (mortality) and 10 and 100 years (time).

(λ) evolves from a negative correlation at the beginning (i.e., the higher the birth rate the lower the number of infected/infectious/recovered) to a positive correlation in the final period (i.e., the higher the birth rate the higher the number of infected/infectious/recovered). From a global population perspective, the highest positive correlation is found with birth rate, whereas the death rate provides a high negative correlation, as expected. The fast progression parameter is also important but, in contrast to the TB-related classes, the correlation in this case is negative.

Female resistance is key to understanding the co-evolution of MtbC and humans. Female protection merits a deeper analysis. Neutralization of both "g" and "s" factors (by giving a value of 1) appears to have not a marked impact in neither the endpoint values (Figs. 3 and 4) nor the sensitivity analysis (Fig. 8 and Suppl. Figure 7). This is because these analyses are based in projections for 100 years. On the contrary, the impact in the "critical" scenarios, i.e. the Paleolithic period infected with "ancient" lineage (Fig. 6) and the Neolithic infected with the "modern" one (Fig. 7), based in projections for 1000 years, is sufficient to change a persistent infection to elimination, together with the host.

Figure 3 shows the impact of these factors on the male/female incidence ratio. It appears that neutralization of the factor "s" (i.e. resistance), which results in equal sex progression to disease, is key as match up the male/female incidence at a proportion of around 50/50. On the contrary, neutralization of tolerance to the disease (g) has a low impact, keeping the ratio from 55/45 to 63/37 depending on the scenario, roughly like with the presence of both protection mechanisms (Std). This analysis precludes that resistance to disease (s) alone is able to explain the difference in incidence based on the sexual dimorphism.

Figure 9 and Suppl. Table 2 illustrates the importance of these parameters in terms of demography in a projection of 100 years. In this figure we analyse the impact of female protection on population growth. The standard simulation (Std) conferred both resistance and tolerance to females by assuming a value of 45/55 for progression to disease (s) and TB mortality μ_{TB} (g), respectively (Std). We subsequently adjusted the birth rate (λ) to ensure

	Paleolithic		Neolithic				
	Ancient	Modern	Ancient	Modern			
	Mortality x 100.000 inhabitants				Scale		
Std	1151	4819	1122	3150	0		
f=0	22	48	3	23	250		
a=0	520	2099	662	1456	500		
w=1	334	468	516	717	1000		
s=1	1428	6129	1172	4246	2000		
g=1	1315	6283	1183	3720	4000		
s&g=1	1679	7875	1234	5922	8000		
	Incidence x 100.000 inhabitants						
Std	4508	18329	3908	10845	0		
f=0	86	186	11	79	812		
a=0	2038	8170	2311	5073	1625		
w=1	1314	1838	1804	2508	3250		
s=1	5693	22620	4181	14444	6500		
g=1	5238	22578	4227	12709	13000		
s&g=1	6612	26907	4389	19182	26000		
	Prevalence x 100.000 inhabitants						
Std	9305	40947	7196	20712	0		
f=0	175	376	21	144	2063		
a=0	4147	17266	4224	9388	4125		
w=1	2669	3741	3295	4594	8250		
s=1	11965	52494	7820	28620	16500		
g=1	10959	52287	7893	24788	33000		
s&g=1	13991	65497	8237	39454	66000		
	Incidence ratio Male/Female					M/F %	
Std	1.47	1.21	1.53	1.36	1.56	61	39
f=0	1.56	1.56	1.54	1.53	1.50	60	40
a=0	1.55	1.36	1.55	1.50	1.38	58	42
w=1	1.48	1.49	1.49	1.49	1.27	56	44
s=1	0.93	0.96	0.96	0.94	1.17	54	46
g=1	1.54	1.21	1.58	1.41	1.08	52	48
s&g=1	1.00	1.00	1.00	1.00	1.00	50	50

Figure 3. End values in the continuous model TBOREX. Influence of progression factors. Heatmap of the projections for 100 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$). Cells are colored according the following rule: minimal (white), and maximal (orange). A color legend has been added per each dataset as a reference.

the viability of the human population under the two "critical" scenarios, namely the Paleolithic and Neolithic with epidemics caused by the "ancient" and "modern" lineages, respectively. Without this protection, the birth ratio should increase from 2.8 to 2.8–3.1 births/fertile female in the Paleolithic and from 4.3–4.6 to 5.2–5.5 in the Neolithic, thus causing an approximate increase of 10% and 20% in birth rates, respectively. When considering only the resistance (s) in females, the impact is the same in the Paleolithic, but decreases to 7% in the Neolithic. Not to mention the impact when comparing with a non-infected population, where the number of births/fertile female could be as low as 2.2 births to keep a steady growth in the Paleolithic and 4 births to have an exponential growth and multiply x20 times the population in 100 years.

"Modern" lineages replace "ancient" ones. Looking at previous results on the incidence of infection, it appears that if a population is infected with both "modern" and "ancient" lineages, the former will predominate due its higher dissemination capacity. We have studied this aspect by building a new model in which we have considered all the coinfection circumstances (Fig. 10). As such, we designed two different scenarios. The simplest scenario initially includes two infectious males, one from each lineage, in a naïve population. The second scenario includes an infectious person carrying a "modern" lineage in the context of a "primed" community where infection with an "ancient" lineage has remained stable after 100 years (Figs. 11 and 12), which seems more realistic considering that "ancient" lineages were the only ones present for a period of more than 20,000 years.

	Paleolithic			Neolithic			Paleolithic			Neolithic		
	Ancient	Modern		Ancient	Modern		Ancient	Modern		Ancient	Modern	
			S end						R end			
Std	2.7	0.0	0.0	43.9	2.2	0	9.9	0.8	0.0	48.4	14.8	0.0
f=0	88.2	52.0	2.8	702.7	562.7	20	1.1	2.4	0.6	0.7	4.1	3.0
a=0	11.4	0.4	5.5	107.1	18.7	40	13.7	6.3	1.2	62.7	44.9	5.5
w=1	13.9	4.4	11.0	115.5	43.2	80	17.1	19.3	2.4	76.1	85.8	11.0
s=1	1.3	0.0	22.0	31.8	0.4	175	8.1	0.2	4.8	42.0	5.1	22.0
g=1	1.7	0.0	44.0	34.3	0.7	350	8.6	0.2	9.5	43.9	7.6	43.0
s&g=1	0.7	0.0	88.0	25.2	0.1	700	6.7	0.0	19.0	37.2	1.5	86.0
			E end						N end			
Std	36.9	1.7	0	233.5	59.8	0	54.5	4.1	0.0	351.1	96.9	0
f=0	23.0	56.1	2	16.3	142.9	10	112.5	110.8	3.5	719.9	710.7	20
a=0	52.1	19.8	4	274.6	204.6	20	80.5	32.0	7.0	463.9	296.0	40
w=1	55.2	53.9	7	292.5	290.4	40	88.6	80.7	14.0	500.6	439.7	80
s=1	27.9	0.4	14	194.5	18.3	80	42.4	1.2	28.0	291.2	33.3	175
g=1	30.5	0.4	28	205.5	29.4	160	45.8	1.2	56.0	308.1	50.2	350
s&g=1	22.0	0.0	56	168.6	4.5	320	34.3	0.2	112.0	251.8	10.0	700
			I end						Clearance			
Std	5.1	1.7	0.0	25.3	20.1	0.0	21.7	11.9	0.0	23.4	13.3	0.0
f=0	0.2	0.4	0.2	0.2	1.0	0.9	77.4	65.7	2.5	82.1	72.2	2.5
a=0	3.3	5.5	0.3	19.6	27.8	1.8	74.2	58.2	5.0	81.1	70.5	5.0
w=1	2.4	3.0	0.7	16.5	20.2	3.5	33.7	13.2	10.0	39.9	26.8	10.0
s=1	5.1	0.6	1.3	22.8	9.5	7.0	21.2	13.8	20.0	26.1	15.8	20.0
g=1	5.0	0.6	2.5	24.3	12.4	14.0	20.1	10.2	40.0	22.4	13.7	40.0
s&g=1	4.8	0.2	5.0	20.7	4.0	28.0	19.1	12.7	80.0	20.8	16.0	80.0

Figure 4. End values and clearance in the TBOREX model simulation. Influence of progression factors. Heatmap of the end values for the projections for 100 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$). Cells are colored according the following rule: minimal **(white)** and maximal (orange). A color legend has been added per each dataset as a reference.

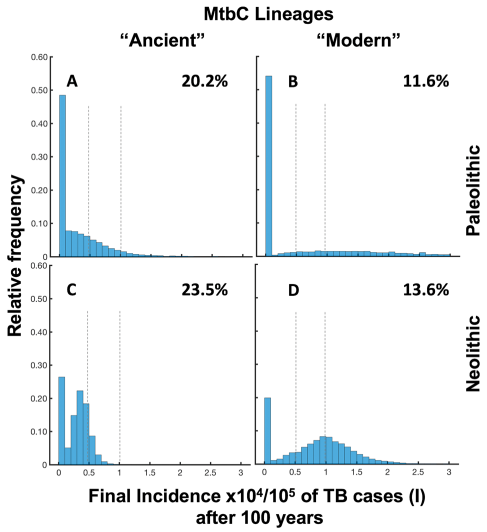

Figure 5. Distribution of final incidence per 100.000 habitants of TB cases in the discrete model TBOREX and infection clearance (absence of EIR). We have run the model 10.000 times. Pictures show the projections for 100 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$). Evolution is drawn under Neolithic life conditions, in relation to the MtbC lineage infection, "Ancient" (**A,C**) or "Modern" (**B,D**). Percentage of clearance is written in the upper right. Dotted lines represent reference values, 5.000 and 10.000 TB cases (**I**)/100.000 h.

The results in the continuous model show how, in naïve communities, "modern" lineages rapidly become predominant in the Paleolithic (Fig. 11), with the "ancient" one disappearing when the population experiences a decline (Suppl. Figure 8). In the case of a "primed" population, "ancient" lineages protect against the entrance of

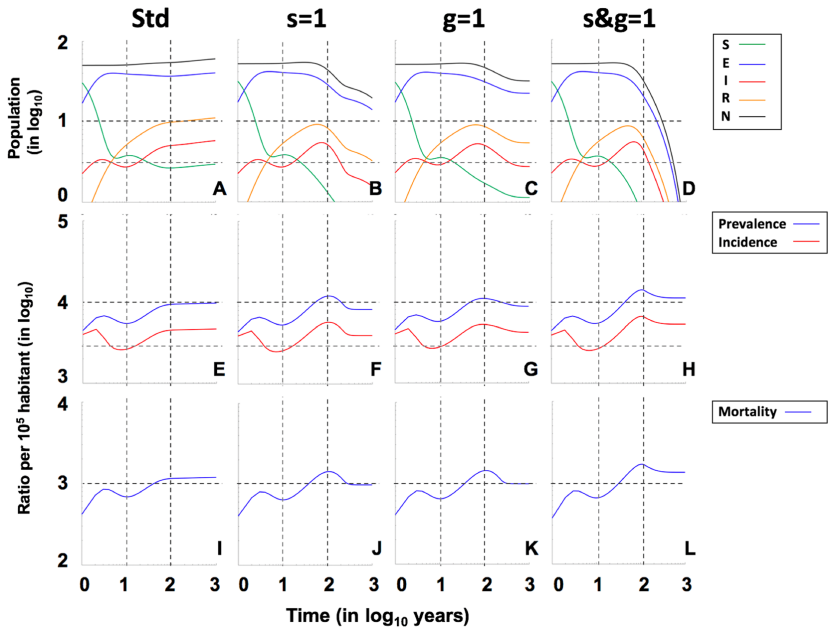

Figure 6. Role of the female tolerance in the evolution of "Ancient" lineages in the Paleolithic. Pictures show the projections for 1000 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$), using continuous resolution of TBOREX model. Evolution is drawn under standard (**Std**) conditions (**A,E,I**), neutralizing the factor"**s**" (**B,F,J**), "**g**" (**C,G,K**) or both, "**s&g**" (**D,H,L**). Continuous lines in pictures **A** to **D** show the total population (**N**) (**black**), Non-infected susceptible (**S**) (**green**), Infected (**E**) (**blue**), TB patients Infectious (**I**) (**red**) and previous TB patients Recovered (**R**) (**orange**). Continuous lines in pictures **E** to **H** show the Incidence (**red**) and Prevalence (**blue**) of TB cases (**I**) per 100.000 inhabitants. Continuous lines in pictures **I** to **L** show the Mortality caused by TB cases per 100.000 inhabitants. Dotted lines represent reference values. We have considered 2 and 10 persons (population), 3.000 and 10.000 (incidence and prevalence), 1.000 (mortality) and 10 and 100 years (time).

"modern" ones, which become predominant 20 years after the appearance thereof. In the Neolithic, the protection of "ancient" lineages lasts for seven years. Both lineages coexist, as we have seen in the Paleolithic, but in this case, thanks to the continuous population growth, the "ancient" lineages do not disappear in any case, even though the "modern" lineage has twice the number of cases (1000). In this case, it is noteworthy the slight presence of some cases (40) of coinfection (I_aI_m) detected in both cases, whereas there is no or only a very residual presence of coinfection under Paleolithic conditions.

Interestingly, when studying the ability of both lineages to persist after analysing the discrete model (Fig. 12), the scenario is different in the case of "primed" populations. In the Paleolithic, "ancient" lineages are able to remain predominant over the "modern" ones (37 vs 23%), with 2% of cases involving both strains and 38% of communities exhibiting global TB clearance. Furthermore, in the Neolithic, the most predominant form after 1000 years is coinfection (76%) followed by the "ancient" lineage (24%), whereas cases with only the "modern" lineage disappear. This scenario changes drastically when the initial conditions are modified. For instance, the entrance of five TB cases with "modern" lineages in a "primed" population leads to a scenario in which the "modern" lineage becomes predominant under Paleolithic conditions (38% of only "modern" to 19% of only "ancient"), whereas 100% of Neolithic communities maintain the presence of both strains after 1000 years (Suppl. Figure 9). These findings are due to two factors: (i) the "ancient" strain is very well established in Neolithic communities, which are larger than Paleolithic ones and therefore prevent the clearance of this strain; and (ii) the possibility of a "modern" strain disappearing during the first years if there is initially only one infectious individual is extremely high when using discrete solving.

It should also be noted that, in this model, we cannot ascertain the impact of these infections on demographics as a logistic growth has been established. As such, this model has been designed simply to evaluate the competition between both MtbC lineages. Indeed, the predominant "modern" lineages had a marked impact in Paleolithic communities, leading to their eradication, and resulted in slow growth with significant mortality in Neolithic ones. The "protective" effect of populations with stabilized "ancient" epidemy, which stopped the progression of "modern" lineages, is noteworthy and can explain their persistence until modern times despite appearing to have lower fitness.

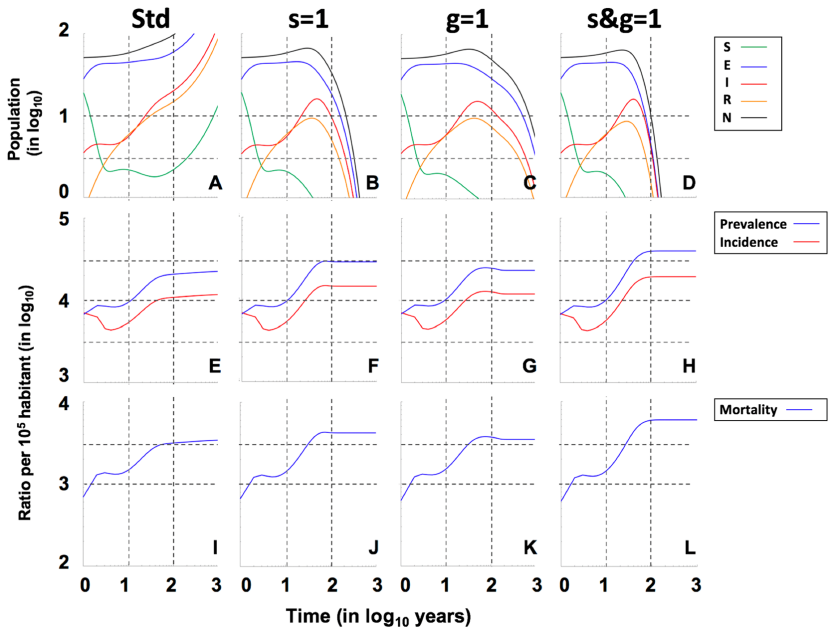

Figure 7. Role of the female tolerance in the evolution of "Modern" lineages in the Neolithic. Pictures show the projections for 1000 years in initial populations of 50 members where a unique male TB patient is included ($I_o = 1$), using continuous resolution of TBOREX model. Evolution is drawn under standard (**Std**) conditions (**A,E,I**), neutralizing the factor"**s**" (**B,F,J**), "**g**" (**C,G,K**) or both, "**s&g**" (**D,H,L**). Continuous lines in pictures **A** to **D** show the total population (**N**) (**black**), Non-infected susceptible (**S**) (**green**), Infected (**E**) (**blue**), TB patients Infectious (**I**) (**red**) and previous TB patients Recovered (**R**) (**orange**). Figure **A** is truncated as the levels grow exponentially to reach final outcomes of >500 persons after 100 years. Continuous lines in pictures **E** to **H** show the Incidence (**red**) and Prevalence (**blue**) of TB cases (**I**) per 100.000 inhabitants. Continuous lines in pictures **I** to **L** show the Mortality caused by TB cases per 100.000 inhabitants. Dotted lines represent reference values. We have considered 2 and 10 persons (population), 3.000 and 10.000 (incidence and prevalence), 1.000 (mortality) and 10 and 100 years (time).

Discussion

Our research supports the notion of a significant increase in population growth before the Neolithic period. This comes from dating of the "modern" MtbC lineages to around 43,000 BCE. Indeed, according to our model, the effect of these strains on humanity was brutal. The entrance of an infectious person with a "modern" strain into a typical Paleolithic human group of hunter-gatherers, i.e. groups of 50 persons with a stable "non-growth" status, resulting in their eradication in around 100 years. However, such groups were not isolated and maintained contacts with other groups for mutual help to hunt large animals, exchange information about new territories to explore, or even to interbreed[10]. This means that the sustainability of the group was not simply dependent on the growth in birth rate and also that the infection spread as a result of contact between tribes, thus resulting in the persistence of MtbC and the subsequent disappearance of these groups. A "Neolithic-like" growth by 43,000 BCE similar to that proposed by Miller *et al.*[23] allows us to understand how "modern" lineages and modern humans were able to persist until modern times.

It should be noted that our results support the proposal that the entrance of "modern" lineages was attenuated by the presence of "ancient" ones. Indeed, the appearance of "ancient" lineages was also a challenge for Paleolithic societies by 73,000 BCE in the context of the most populated continent (Africa) and the different migrations towards the Levant[9]. In order to sustain these societies, we considered a growth rate that could double the population in 100 years in the case of being free from TB, in other words a growth of around 1%/year, which is closer to the Contemporary history growth. This growth is not supported by any previous studies, with reported values not exceeding 0.003%/year[13,52], unless we include the impact of TB itself on natural mortality. This impact should lead to further research in this field given that a majority of humans were living with TB in the Paleolithic. In fact, after analysing the data, we should maybe have changed change the term "Neolithic" for "Neolithic-like" conditions, although we decided not to do so for the sake of clarity.

One of our key innovations has been to consider the birth rate to be the main source of population growth. This variable is missing in in previous models[28–30], which simplify it by assuming a logistic growth up to a certain level in order to achieve stabilization of the epidemics. By including this variable, we wanted to understand

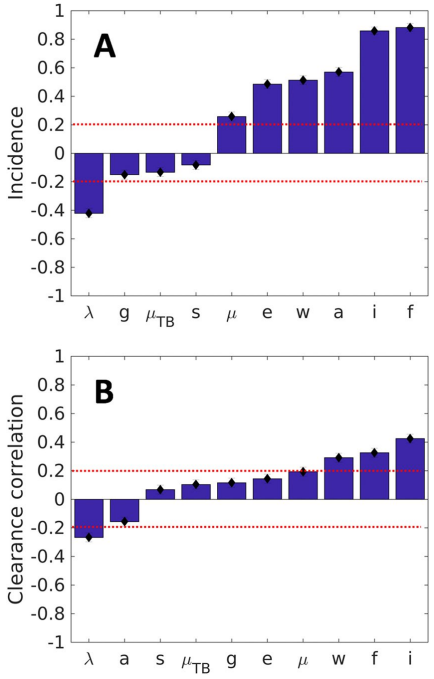

Figure 8. Sensitivity analysis. Partial Rank Correlation Coefficient on the TBOREX discrete model analyzing 1000 simulations. Parameters: birth rate (λ), natural mortality (μ), mortality caused by TB (μ_{TB}), fast progression (f) and reactivation factor (a), number of infections caused by a patient (e), reactivation factor in recovered (w), tolerance to progression to disease (s) and to disease (g), and immunity (i). Picture A shows the influence of each parameter in the incidence and B clearance of the infection. Dotted red lines show the variation of a 20%.

Parameter	Paleolithic value	Neolithic value	Sensitivity analysis range
λ	0.078	0.128	[0.07 0.13]
μ	0.03030	0.03846	[0.0286 0.04]
μ_{TB}	0.12	0.15	[0.1 0.17]
e	10	20	[10 20]
f	0.099(A)/0.0825(M)	0.1238(A)/0.1031(M)	[0 0.13]
i	0.1	0.1	[0.05 0.5]
w	7	7	[1 7]
a	0.0297(A)/0.02475(M)	0.03714(A)/0.03093(M)	[0 0.038]
g	1.2222	1.2222	[1 1.4]
s	1.2222	1.2222	[1 1.4]

Table 2. Sensitivity analysis.

the impact of TB on demographics, and we have been able to address the question of the lower TB incidence in females. This opens the way to ascertaining the biological mechanisms that make females less susceptible to both TB and other severe epidemics and famines, as has been noted recently[53]. It also opens up eco immunological concepts from the general trade-off between immunity and fertility based on Bateman's principle of immunity or the "immunocompetence handicap hypothesis"[54,55] towards a sex-specific investment in defense, depending on the nature of the infecting parasite[56,57].

That is why we have refined our model assuming a female resistance and tolerance to progression of the infection (s) and to the disease itself (g), respectively, in order to reflect a different incidence between males and

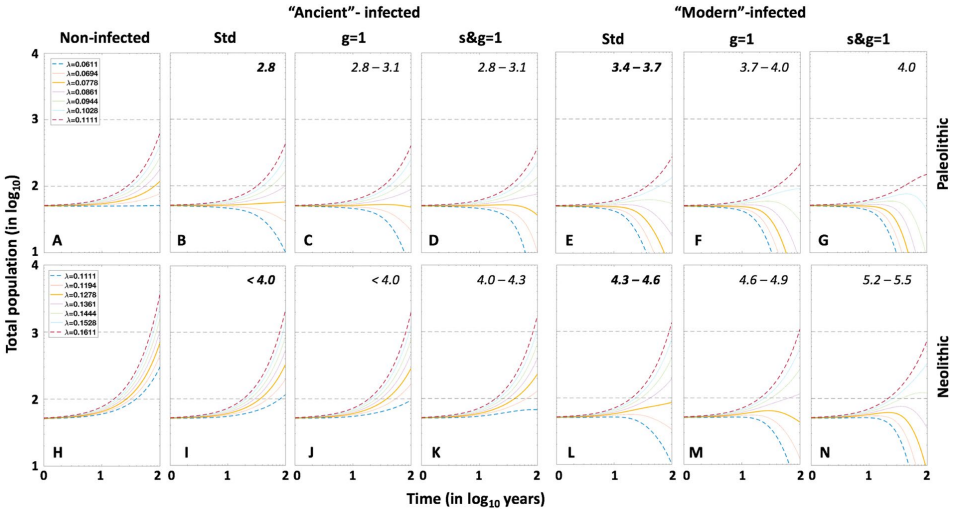

Figure 9. Natality and the evolution of total population. Influence of sex and strain. Evolution of total population, using continuous resolution of TBOREX model, with different natality indexes in TB-free communities (Non-infected), standard conditions (Std), neutralizing factor g (g = 1) and neutralizing factors g and s (s&g = 1). Continuous lines show the dynamics with the natality values used in the model (0.0778 in Paleolithic conditions and 0.1278 in Neolithic).

females. Indeed, the concept of a balanced response to avoid damage to the host, thus obtaining a better fitness, has been included in the study of human infectious diseases by Casadevall and Pirofski[58], who clearly showed that an exaggerated response against MtbC is one of the mechanisms for developing active TB. In fact, it is the main factor. Immunosuppression, which can be exemplified by HIV infection as the most extreme and widely distributed example, causes no more than 10% of TB cases worldwide[3]. This knowledge, which has recently been applied to the TB field, has led to a new prophylactic and therapeutic era focused on "host-derived therapies"[59].

Our data suggest that female resistance predominates over tolerance. This is supported by the data obtained in the Neolithic scenario, which show that only an increase in resistance can explain the reduced incidence in females. The explanation for this resistance could lie in the enhanced level of Tregs generated by oestrogens. As demonstrated in a non-human primate experimental model, latently infected animals with increased levels of Tregs exhibit less progress to active TB[60]. The mechanism is based on the ability of Tregs to reduce local inflammatory responses, especially neutrophil infiltration, which fuels the growth of Mtb[61]. Interestingly, the Tregs-induced tolerance mechanism may be responsible for reducing the bacillary load, as explained in detail previously[62], thus generating resistance[35]. In this case, for simplicity, we have allocated this property to the ability to develop TB disease, but with a lower mortality. The problem is that the development of TB in females may cause a loss of global human fitness by reducing the reproductive capacity. TB in women generates more extrapulmonary forms than in males, thus resulting in a decrease in fertility of up to 40%. It also increases the perinatal mortality six fold and is the cause of 6–10% of all maternal deaths[63–65]. For simplicity, these data have not been included in our model, which nevertheless supports the concept that a resistance mechanism to avoid the development of TB in women is more likely to explain the co-evolution of humans and MtbC, especially when taking into account the extraordinary consequences on demographics.

One of the main factors affecting the birth rate in the Paleolithic identified by experts is the late weaning of children and the resulting contraceptive effect of breastfeeding[26]. It is still unclear, however, why weaning was brought forward in the agricultural period, thus increasing the birth rate. Although modesty is currently the only explanation, one can speculate the creation of a matriarchal society that can exert demographic control. Indeed, the concept of self-sustainability also appears to be present[10]. As such, we can speculate that TB epidemics stimulated this increase in birth rate, especially after the appearance of "modern" lineages, in a way that selected the communities based on it, thus allowing those that adopted early weaning to persist. Communities with a lower birth rate disappeared.

Our work also has incorporated other assumptions regarding the different virulence mechanisms between "ancient" and "modern" MtbC lineages in the knowledge that this is a question open to discussion. What is clear is the evidence for the higher dissemination ability of "modern" lineages compared with "ancient" ones[21]. The fact that the latter are restricted to specific geographical regions also points to some sort of genetic susceptibility linked to race/ethnic groups, as proposed by several authors[7,17,66]. Our model has not considered this aspect. However, we have established a better ability to infect contacts (e) in "modern" lineages that is twice as high as in "ancient" ones, in the range previously established by Styblo[67]. We have also incorporated a higher fast progression (f) for

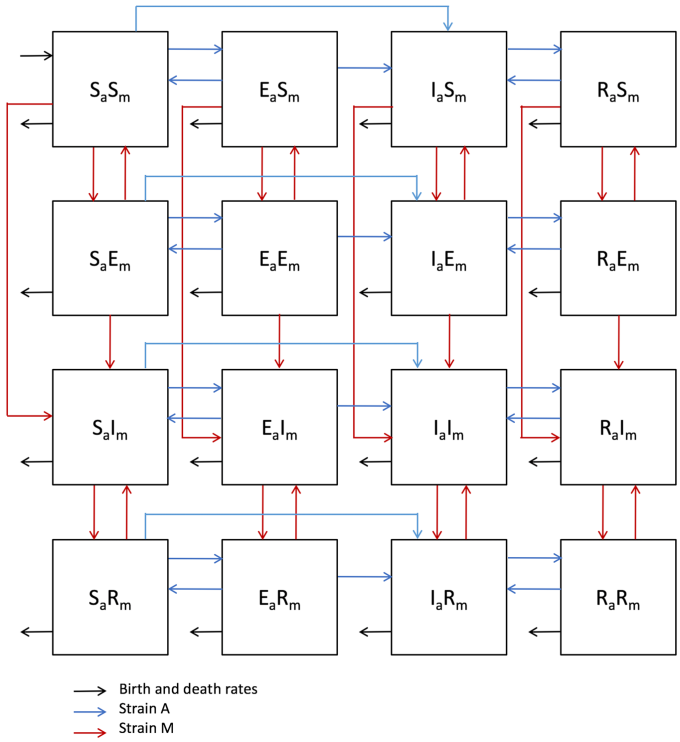

Figure 10. The Coinfection model. Each compartment refers to the set of individuals by disease status with regards to "ancient" (sub index a) and "modern" (sub index m) strains: Susceptible (**S**), Exposed (**E**), Infected (**I**), Recovered (**R**). New-born individuals are assumed susceptible. Color of arrows and subheadings refers to the evolution of the infection of each strain, corresponding to "ancient" lineage (**a**) (**blue**) and "modern" lineage (m) (**red**). The evolution of the coinfection causes not only the corresponding EIR population but also coinfection, where $\mathbf{E_aE_m}$ represents a person with a latent coinfection, $\mathbf{I_aE_m}$ an Infectious person with an "ancient" strain, latently infected with a "modern" strain, $\mathbf{E_aI_m}$ the reverse, etc.

"ancient" lineages based on their higher pro-inflammatory properties[16], assuming the limitation of the range of strains used. In this regard, we have hypothesized that "modern" lineages lost the pro-inflammatory ability of "ancient" lineages by decreasing their bacillary size[19], thus favouring the induction of smaller aerosols and acquiring a better fitness in terms of ability to disseminate.

The uncertainty analysis clearly shows that fast progression (f) is the most important factor as regards increasing the incidence but is also responsible for accelerating clearance. In this regard, a high transmission capacity (e) has a more balanced role by promoting the incidence but having less influence on clearance of the infection in the community. Thus, we can easily explain the success of "modern" lineages even though our results give a predominant role to the reactivation factor, which is higher for "ancient" lineages. Reactivation is an important factor for increasing the incidence while at the same time preventing clearance of the infection. That is why we consider that the balance of both factors (f and a) is very important as regards allowing both the persistence of MtbC and humankind. These findings complement previous hypotheses concerning the relevance of endogenous reactivation when it comes to understanding the persistence of MtbC[27,28], but refines them by considering reliable experimental data and taking into consideration that, generally speaking, the MtbC in the Paleolithic was quite similar to that found nowadays. This is consistent with the paradigmatic genomic stability found in MtbC[68]. Our analysis also reveals the importance of having a good immunity, thus having the lowest possible value of "i". In this model, we have used a conservative proxy of the value obtained from the observations reported by Heimbock[69], which have recently been emphasized by Bloom[70], concerning the protection triggered by natural MtbC infection against developing active TB (about 97%), and the need to obtain a prophylactic vaccine that can increase this protection while avoiding the risk of fast progression (f) and endogenous reactivation (a).

It is interesting to note the significant effort invested to build a coinfection model applicable to both lineages and to validate the ability of "modern" lineages to replace "ancient" ones in conditions of coinfection. In fact, this

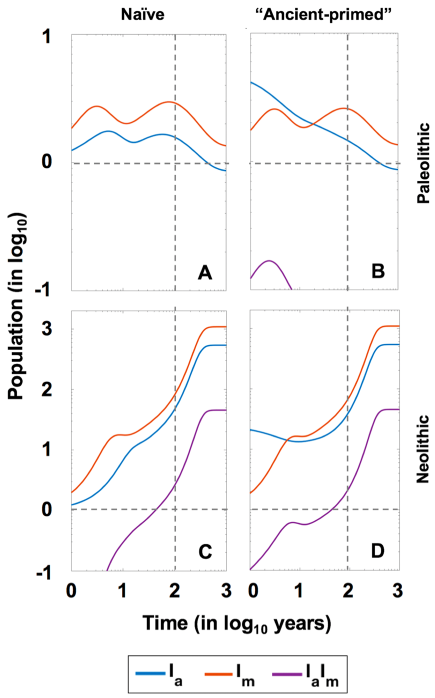

Figure 11. The replacement of "ancient" lineages by "modern" ones in a continuous coinfection model. Pictures show the evolution of the coinfection in four different circumstances, using the continuous coinfection model. Left quadrants show the evolution of the coinfection in the context of a "naive" population with two infectious males, one from each lineage, in the periods of Paleolithic (**A**) and Neolithic (**C**). Right quadrants show the evolution of the coinfection in the context of a community where infection with an "ancient" lineage has remained stable for 100 years and one "modern" lineage infected male is introduced, in the periods of Paleolithic (**B**) and Neolithic (**D**).

can be characterized as a superinfection model as a more virulent strain infects a previously infected host. The disadvantage of more virulent strains is that they kill the host faster, thus causing local extinction of the hosts[71]. The paradox of our model is that we are faced with a lineage that is better able to progress to disease, and thus be transmissible (the "ancient"), and another one with an ability to cause disease but with a greater ability to disseminate (the "modern"). To the best of our knowledge, this is the first attempt to do so. Even when applying a logistic limit to population growth, we have been unable to discern the demographic impact of coinfection, although the model corroborates the higher dissemination of "modern" lineages as a winning strategy while illustrating the protective role of communities previously infected with "ancient" ones. In this regard, the induction of immunity, and the reduction in the susceptible population, explains this phenomenon. It also explains how "ancient" lineages have been able to persist when faced with such competition, becoming limited geographically[21]. In this regard, the incidence of "ancient" covers the territories implicated in the second "out of Africa" migration, from West Africa towards India and Australia[9].

In summary, our model shows the marked impact of TB on human history, from the Paleolithic, a fact that should be revisited and included in future studies to interpret this human history. Our model agrees with the more recent data showing a demographic explosion prior to the Neolithic revolution. We also consider factors such as the higher vulnerability of recovered patients which, especially under current conditions, still represent 40–50% of undiagnosed TB cases in Asia and Africa. Finally, we highlight the importance of female resistance to understand the lower incidence in women, a fact that should be studied in order to discern its biological basis.

Methodology

Basic TB natural history compartmental model. We have designed a compartmental mathematical model based on five differential equations to describe the dynamics of the evolution of MtbC infection in the population, based on previous models[28–30,51] (Supplementary Figure 1 and Table 1). We have classified the whole population (N) into compartments following the classical SEIR [susceptible (S), exposed/infected (E), infectious

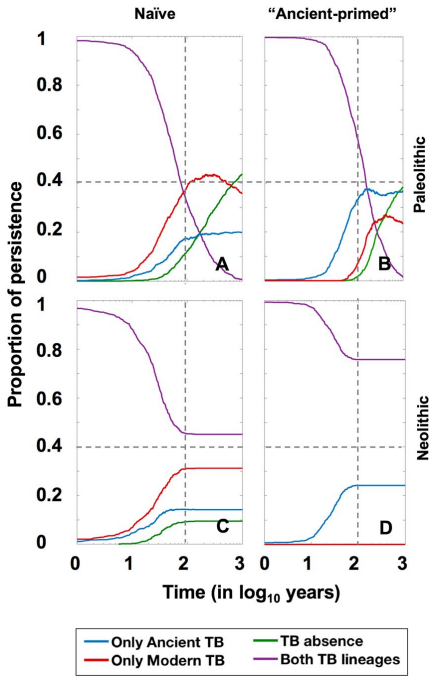

Figure 12. The replacement of "ancient" lineages by "modern" ones in a discrete coinfection model. Pictures show the persistence of the coinfection in four different circumstances, using the discrete coinfection model. 10000 independent repetitions of each scenario are run and the proportion of communities with presence of both strains (purple), only ancient (blue), only modern (red) or none of them (green) along the time are shown. Left quadrants show the persistence in the context of a "naive" population with two infectious males, one from each lineage, in the periods of Paleolithic (**A**) and Neolithic (**C**). Right quadrants show the persistence in the context of a community where infection with an "ancient" lineage has remained stable for 100 years and one "modern" lineage infected male is introduced, in the periods of Paleolithic (**B**) and Neolithic (**D**).

(I) and recovered (R)] approach, where $N = S + E + I + R$. The evolution of these five variables is described by Eqs. (1–5):

$$\frac{dS}{dt} = \pi + \delta E - \mu S - \beta IS \tag{1}$$

$$\frac{dE}{dt} = (1 - f)\,\beta SI\ - (\mu + \delta + i(a + rI))E \tag{2}$$

$$\frac{dI}{dt} = f\beta SI + i(a + rI)E + wi(a + rI)R - (\mu + \mu_{TB} + c)I \tag{3}$$

$$\frac{dR}{dt} = cI - (\mu + wi(a + rI))R \tag{4}$$

$$\frac{dN}{dt} = \frac{dS}{dt} + \frac{dE}{dt} + \frac{dI}{dt} + \frac{dR}{dt} \tag{5}$$

Infected aerosols are released by infectious subjects (I) and can infect susceptible individuals (S) or reinfect already infected ones (E and R), with an annual risk of infection (β) that depends on the number of new infections (e) caused by a particular case (I). This is related to the growth ratio of the population, and thus the birth rate (π) and natural mortality (μ). Thus, the annual risk of infection is described by the relationship $\beta = e \cdot \pi/\mu$[30].

After infection or reinfection, subjects can develop the disease during the first year according to the probability f, known as fast progression. This probability has been taken from the most recent studies concerning progression from infection to disease[45,72]. The most detailed of these studies[45] gives a value of 8.25% for the whole population. Furthermore, once an infected (E) or recovered status (R) has been achieved, there is a chance of reactivation during the following four years, which according to these authors represents 1.5% of the fast progression (f) and is termed the reactivation factor (a). In light of this, we have defined the risk of disease caused by reinfection (r) as $r = f\beta + a(1 - f)\beta$. This value is substantially decreased by the immunity (i) generated by infection. Considering the studies of Heimbeck[69], those infected people that do not develop the disease have a protection against the onset of disease of at least 90%, thus resulting in a protective ratio (i) of 0.1.

Infected people (E) can both drain the bacillary load and reduce the immunity (δ), depending on the dynamic nature of the infection[41] during the period of around 10 years established in BCG immunity studies[40]. This would lead to a drainage probability of about 0.1 annually, although this is reduced by the possibility of endogenous or exogenous reactivation of the infection, defined as a and r, respectively. Thus, bacillary drainage is defined as: $\delta = 0.1 - a - r$.

According to Tiemerman *et al.*[47], TB patients can cure naturally in three years, thus giving an annual curation rate (c) of 0.33. They also have a global chance of dying of 45% during these three years, therefore the annual mortality caused by TB (μ_{TB}) is 0.15. Furthermore, Uys *et al.*[31] have determined that recovered subjects have a sevenfold higher chance of developing disease, thus we have included this factor (w) in our model. The birth rate (π) is determined from the mean number of births per fertile woman/year (λ).

In light of the previous findings of Bold *et al.* and Portevin *et al.*[16,19], we assumed that, given their ability to induce a higher proinflammatory response to macrophages, "ancient" lineages were able to increase the probability of fast progression (f) and reactivation (a) more than their "modern" counterparts. The percentage increase resulting from the higher inflammation was established as 20% as this is the relative volume of the upper lobes, with this property being determinant for the onset of disease[18]. We consider that the qualitative advantage of "modern" lineages arises due to the reduction of the size of the bacilli[19]. As such, "modern" lineages replace the "ancient" pro-inflammatory advantage by developing more infective aerosols, thus substantially increasing the number of infections caused by a patient (e). Given the findings of Styblo[67], we have situated these values as the average in the case of "ancient" lineages and the upper limit for the "modern" ones, in other words a value of 10 and 20, respectively.

Our model considers the concepts of resistance to infection, which limits the progression to disease (i.e. reducing the bacillary load), and tolerance to disease, which limits the mortality caused by TB (μ_{TB}). This is because tolerance reduces the damage caused by MtbC infection and increases the fitness of the host[58]. In order to quantify these differences in general mortality between the Paleolithic and Neolithic periods, we have considered the difference in life expectancy, estimated as 33 and 26.5 years, respectively[73,74]. This 25% difference has been taken into account to determine the change in resistance, thus affecting the fast progression (f) and reactivation factor (a), and changes in tolerance, thus impacting TB-related mortality (μ_{TB}).

The reader can find a figure with the model's flow chart and a table with the parameters' values in the article (Fig. 1A and Table 1).

TBOREX model. In order to study the observed sex-related variations in incidence, we have refined the basic model by building the TBOREX (TB, Origen and Sex) one, which take into consideration male (M) and female (F) subpopulations in each compartment [susceptible (S), exposed/infected (E), infectious (I) and recovered (R)]. This results in the following differential Eqs. (6–15),

$$\frac{dS_M}{dt} = \pi + \delta E_M - \mu S_M - \beta(I_M + I_F)S_M \tag{6}$$

$$\frac{dE_M}{dt} = (1 - fs)\beta(I_M + I_F)S_M - (\mu + \delta + is(a + r(I_M + I_F)))E_M \tag{7}$$

$$\frac{dI_M}{dt} = sf\beta(I_M + I_F)S_M + is(a + r(I_M + I_F))E_M + wis(a + r(I_M + I_F))R_M - (\mu + g\mu_{TB} + c)I_M \tag{8}$$

$$\frac{dR_M}{dt} = cI_M - (\mu + wis(a + r(I_M + I_F)))R_M \tag{9}$$

$$\frac{dN_M}{dt} = \frac{dS_M}{dt} + \frac{dE_M}{dt} + \frac{dI_M}{dt} + \frac{dR_M}{dt} \tag{10}$$

$$\frac{dS_F}{dt} = \pi + \delta E_F - \mu S_F - \beta(I_M + I_F)S_F \tag{11}$$

$$\frac{dE_F}{dt} = \left(1 - \frac{1}{s}f\right)\beta(I_M + I_F)S_F - \left(\mu + \delta + i\frac{1}{s}(a + r(I_M + I_F))\right)E_F \tag{12}$$

$$\frac{dI_F}{dt} = \frac{1}{s}f\beta(I_M + I_F)S_F + i\frac{1}{s}(a + r(I_M + I_F))E_F + wi\frac{1}{s}(a + r(I_M + I_F))R_F - \left(\mu + \frac{1}{g}\mu_{TB} + c\right)I_F \tag{13}$$

$$\frac{dR_F}{dt} = cI_F - \left(\mu + wi\frac{1}{s}(a + r(I_M + I_F))\right)R_F \tag{14}$$

$$\frac{dN_F}{dt} = \frac{dS_F}{dt} + \frac{dE_F}{dt} + \frac{dI_F}{dt} + \frac{dR_F}{dt} \tag{15}$$

The total population values in each compartment can be evaluated simply by adding both subpopulations (e.g., $N = N_M + N_F$).

There is a wide consensus regarding the stability of the population density in the Middle Paleolithic. Thus, it appears that the growth rate was around 0.003%, which significantly increased up to the 0.1% in the Neolithic[13,26,52]. We have adjusted this rate for the expected predominant lineage in each period, namely the "ancient" one for the Paleolithic and the "modern" for the Neolithic, to determine the birth rate (λ). This gave a value of $\lambda = 0.0778$, which means an average of 2.8 children/female in the Paleolithic period. For the Neolithic, we considered a value of $\lambda = 0.128$, thus meaning an average of 4.7 children/female. In both cases we considered that 50% of females are fertile every year, with a fertility period of 18 years starting at the age of 15. These fertility values give a much higher population growth than that determined by different experts in this field[13,14,26,52]. The birth rate in each susceptible compartment (susceptible male and susceptible female) can be represented as $\pi = 0.5 \cdot \lambda \cdot N_F$, where N_F represents the total number of females. In order to maintain the actual female/male incidence proportion at the accepted value of 60/40, we adjusted the progression to disease (resistance) by a factor (s) that increases or reduces this in men and women, respectively, in a proportion of 55/45. We also added this proportion to the mortality caused by TB (μ_{TB}) in order to reproduce tolerance to disease (g).

The reader can find a figure with the model's flow chart and a table with the parameters' values in the article (Fig. 1B and Table 1).

Assessment of uncertainty and sensitivity in the system. An uncertainty and sensitivity analysis were performed for the TBOREX model as described in[75] using a sampling-based method. Thus, 1000 different parameter sets were used to explore the space using a Latin Hipercube Sampling (LHS) technique. Parameters were explored between the values shown in Table 2. At each time step, the Partial Rank Correlation Coefficient (PRCC) was computed for each of the parameters and susceptible, exposed, infected, recovered and total populations, as well as the annual incidence and death rate. The final PRCC between input parameters and TB clearance was also computed, using the discrete resolution. This methodology allowed us to see how each output was affected upon increasing (or decreasing) a specific parameter (linearly discounting the effects of the uncertainty on the rest of the parameters). Thus, PRCC can be used to determine which parameters to target to achieve specific goals.

Coinfection model. The coinfection model takes into consideration the coexistence of "ancient" (a) and "modern" (m) lineages in a certain community. The variables of the model are the susceptible individuals (S), the exposed (E), the infectious (I) and the recovered (R). E_aS_m, S_aE_m and E_aE_m represent persons with an ancient, a modern or both latent infections, respectively. The same nomenclature is applied to all possible combinations between S_a, E_a, I_a and R_a – i.e., TB compartments related with ancient strain – and S_m, E_m, I_m and R_m – i.e., TB compartments related with modern strain. For simplicity, in these equations we will use: $I_a = I_aS_m + I_aE_m + I_aR_m$ (i.e., all the compartments with ancient strain TB-sick); $I_m = S_aI_m + E_aI_m + R_aI_m$ (i.e., all the compartments with modern strain TB-sick); $I_{am} = I_aI_m$ (i.e., the compartment with TB-sick with the two strains). The evolution of all variables is described by Eqs. (16–31):

$$\frac{d}{dt}S_aS_m = \Pi - \beta_a \cdot S_aS_m \cdot I_a - \beta_m \cdot S_aS_m \cdot I_m + \delta_m \cdot S_aE_m + \delta_a \cdot E_aS_m - \mu \cdot S_aS_m \tag{16}$$

$$\frac{d}{dt}E_aS_m = \delta_m \cdot E_aE_m - i(a_a + r_aI_a) \cdot E_aS_m + (1 - f_a)\beta_a \cdot S_aS_m \cdot I_a - \beta_m \cdot E_aS_m \cdot I_m - \delta_a \cdot E_aS_m + \delta_m \cdot E_aE_m - \mu \cdot E_aS_m \tag{17}$$

$$\frac{d}{dt}I_aS_m = \delta_m \cdot I_aE_m - c \cdot I_aS_m + wi(a_a + r_aI_a) \cdot R_aS_m + i(a_a + r_aI_a) \cdot E_aS_m + f_a\beta_a \cdot S_aS_m \cdot I_a - \beta_m \cdot S_aS_m \cdot I_m - (\mu + \mu_{TB}) \cdot I_aS_m \tag{18}$$

$$\frac{d}{dt}R_aS_m = \delta_m \cdot R_aE_m - wi(a_a + r_aI_a) \cdot R_aS_m + c \cdot I_aS_m - \beta_m \cdot S_aS_m \cdot I_m - \mu \cdot R_aS_m \tag{19}$$

$$\frac{d}{dt}S_aE_m = \delta_a \cdot E_aE_m - \delta_m \cdot S_aE_m - i(a_m + r_mI_m) \cdot S_aE_m - \beta_a \cdot S_aS_m \cdot I_a + (1 - f_m)\beta_m \cdot S_aS_m \cdot I_m - \mu \cdot S_aE_m \tag{20}$$

$$\frac{d}{dt}E_aE_m = -\delta_a \cdot E_aE_m - \delta_m \cdot E_aE_m - i(a_a + r_aI_a) \cdot E_aE_m - i(a_m + r_mI_m) \cdot E_aE_m + (1 - f_a)\beta_a \cdot S_aE_m \cdot I_a + (1 - f_m)\beta_m \cdot E_aS_m \cdot I_m - \mu \cdot E_aE_m \tag{21}$$

$$\frac{d}{dt}I_aE_m = -\delta_m \cdot I_aE_m + wi(a_a + r_aI_a) \cdot R_aE_m - c \cdot I_aE_m + i(a_a + r_aI_a) \cdot E_aE_m - i(a_m + r_mI_m) \cdot I_aE_m + f_a\beta_a \cdot S_aE_m \cdot I_a + (1 - f_m)\beta_m \cdot I_aS_m \cdot I_m - (\mu + \mu_{TB}) \cdot I_aE_m \tag{22}$$

$$\frac{d}{dt}R_aE_m = -\delta_m \cdot R_aE_m - wi(a_a + r_aI_a) \cdot R_aE_m + c \cdot I_aE_m - i(a_m + r_mI_m) \cdot R_aE_m + (1 - f_m)\beta_m \cdot R_aS_m \cdot I_m - \mu \cdot R_aE_m \tag{23}$$

$$\frac{d}{dt}S_aI_m = \delta_a \cdot E_aI_m + wi(a_m + r_mI_m) \cdot S_aR_m - c \cdot S_aI_m + i(a_m + r_mI_m) \cdot S_aE_m - \beta_a \cdot S_aS_m \cdot I_a + f_m\beta_m \cdot S_aS_m \cdot I_m - (\mu + \mu_{TB}) \cdot S_aI_m \tag{24}$$

$$\frac{d}{dt}E_aI_m = -\delta_a \cdot E_aI_m + wi(a_m + r_mI_m) \cdot E_aR_m - c \cdot E_aI_m - i(a_a + r_aI_a) \cdot E_aI_m + i(a_a + r_aI_a) \cdot E_aE_m + (1 - f_a)\beta_a \cdot S_aI_m \cdot I_a + f_m\beta_m \cdot E_aS_m \cdot I_m - (\mu + \mu_{TB}) \cdot E_aI_m \tag{25}$$

$$\frac{d}{dt}I_aI_m = wi(a_m + r_mI_m) \cdot I_aR_m + wi(a_a + r_aI_a) \cdot R_aI_m - 2c \cdot I_aI_m + i(a_a + r_aI_a) \cdot E_aI_m + i(a_m + r_mI_m) \cdot I_aE_m + f_a\beta_a \cdot S_aI_m \cdot I_a + f_m\beta_m \cdot I_aS_m \cdot I_m - (\mu + 2\mu_{TB}) \cdot I_aI_m \tag{26}$$

$$\frac{d}{dt}R_aI_m = wi(a_m + r_mI_m) \cdot R_aR_m - wi(a_a + r_aI_a) \cdot R_aI_m + c \cdot I_aI_m - c \cdot R_aI_m + i(a_m + r_mI_m) \cdot R_aE_m + f_m\beta_m \cdot R_aS_m \cdot I_m - (\mu + \mu_{TB}) \cdot R_aI_m \tag{27}$$

$$\frac{d}{dt}S_aR_m = \delta_a \cdot E_aR_m - wi(a_m + r_mI_m) \cdot S_aR_m + c \cdot S_aI_m - \beta_a \cdot S_aS_m \cdot I_a - \mu \cdot S_aR_m \tag{28}$$

$$\frac{d}{dt}E_aR_m = -\delta_a \cdot E_aR_m - wi(a_m + r_mI_m) \cdot E_aR_m + c \cdot E_aI_m - i(a_a + r_aI_a) \cdot E_aR_m + (1 - f_a)\beta_a \cdot S_aR_m \cdot I_a - \mu \cdot E_aR_m \tag{29}$$

$$\frac{d}{dt}I_aR_m = -wi(a_m + r_mI_m) \cdot I_aR_m + wi(a_a + r_aI_a) \cdot R_aR_m - c \cdot I_aR_m + c \cdot I_aI_m + i(a_a + r_aI_a) \cdot E_aR_m + f_a\beta_a \cdot S_aR_m \cdot I_a - (\mu + \mu_{TB}) \cdot I_aR_m \tag{30}$$

$$\frac{d}{dt}R_aR_m = -\ wi(a_m + r_mI_m) \cdot R_aR_m - wi(a_a + r_aI_a) \cdot R_aR_m + c \cdot I_aR_m + c \cdot R_aI_m - \mu \cdot R_aR_m \quad (31)$$

where:

$r_a = f_a\beta_a + a_a(1 - f_a)\beta_a r_m = f_m\beta_m + a_m(1 - f_m)\beta_m$
$\delta_a = 0.1 - a_a - r_a\,\delta_m = 0.1 - a_m - r_m$
$\beta_a = e_a\mu/\pi\ \beta_m = e_m\mu/\pi$

The reader can find a figure with the model's flow chart and a table with the parameters' values in the article (Fig. 2 and Table 1).

Continuous and discrete resolution of the models. The two models were numerically integrated with Matlab using the Euler method, with an integration step of 1/36 years. This resulted in curves showing the evolution of each variable, as well as the annual incidence and mortality. The models were also resolved using Matlab's ode45 package (MathWorks, Natick, Massachusetts, EEUU) to verify the correctness of the manual resolution.

The limited size of some of the communities studied suggests the suitability of exploring a discrete resolution of the models, using natural numbers to describe the variable dynamics. To that end, discrete resolution was implemented using Euler's integration method but converting each of the flows at each integration step into a natural number using Poisson random distribution. As such, discrete resolution of the models provides a different solution due to the effect of randomness.

Received: 12 October 2019; Accepted: 9 December 2019;
Published online: 08 January 2020

References

1. Paulson, T. Epidemiology: A mortal foe. Nature 502, S2-3 (2013).
2. WHO. TB: a global emergency, WHO report on the TB epidemic. (1994).
3. WHO. Global Tuberculosis Report 2018. (2018).
4. Stead, W. W. The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin. Chest Med.* **18**, 65–77 (1997).
5. Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc. Natl. Acad. Sci.* **99**, 3684–3689 (2002).
6. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* **45**, 1176–82 (2013).
7. Brites, D. & Gagneux, S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. *Immunol. Rev.* **264**, 6–24 (2015).
8. Chisholm, R. H., Trauer, J. M., Curnoe, D. & Tanaka, M. M. Controlled fire use in early humans might have triggered the evolutionary emergence of tuberculosis. *Proc. Natl. Acad. Sci. USA* **113**, 9051–6 (2016).
9. Oppenheimer, S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 770–84 (2012).
10. Armelagos, G. J., Goodman, A. H. & Jacobs, K. H. The origins of agriculture: Population growth during a period of declining health. *Popul. Environ.* **13**, 9–22 (1991).
11. Johnson, A. W. & Earle, T. K. The evolution of human societies: from foraging group to agrarian state. (Stanford University Press, 1987).
12. Gurven, M. & Kaplan, H. Longevity Among Hunter- Gatherers: A Cross-Cultural Examination. *Popul. Dev. Rev.* **33**, 321–365 (2007).
13. Cohen, M. N. The food crisis in prehistory: overpopulation and the origins of agriculture. (Yale University Press, 1977).
14. Zahid, H. J., Robinson, E. & Kelly, R. L. Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proc. Natl. Acad. Sci.* **113**, 931–935 (2016).
15. Guzmán, R. A. & Weisdorf, J. The Neolithic Revolution from a price-theoretic perspective. *J. Dev. Econ.* **96**, 209–219 (2011).
16. Portevin, D., Gagneux, S., Comas, I. & Young, D. Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. *PLoS Pathog.* **7**, e1001307 (2011).
17. Saelens, J. W., Viswanathan, G. & Tobin, D. M. Mycobacterial Evolution Intersects With Host Tolerance. *Front. Immunol.* **10**, 528 (2019).
18. Cardona, P.-J. The key role of exudative lesions and their encapsulation: lessons learned from the pathology of human pulmonary tuberculosis. *Front. Microbiol.* **6**, 612 (2015).
19. Bold, T. D. *et al.* Impaired fitness of Mycobacterium africanum despite secretion of ESAT-6. *J. Infect. Dis.* **205**, 984–90 (2012).
20. Fennelly, K. P. & Jones-López, E. C. Quantity and Quality of Inhaled Dose Predicts Immunopathology in Tuberculosis. *Front. Immunol.* **6**, 313 (2015).
21. Wiens, K. E. *et al.* Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Med.* **16**, 196 (2018).
22. Atkinson, Q. D., Gray, R. D. & Drummond, A. J. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol. Biol. Evol.* **25**, 468–74 (2008).
23. Miller, E. F., Manica, A. & Amos, W. Global demographic history of human populations inferred from whole mitochondrial genomes. *R. Soc. open Sci.* **5**, 180543 (2018).
24. Price, T. D. & Bar-Yosef, O. The Origins of Agriculture: New Data, New Ideas. *Curr. Anthropol.* **52**, S163–S174 (2011).
25. Leonardi, M., Barbujani, G. & Manica, A. An earlier revolution: genetic and genomic analyses reveal pre-existing cultural differences leading to Neolithization. *Sci. Rep.* **7**, 3525 (2017).
26. Gilligan, I. *Climate, Clothing, and Agriculture in Prehistory*, https://doi.org/10.1017/9781108555883 (Cambridge University Press, 2018).
27. Blaser, M. J. & Kirschner, D. The equilibria that allow bacterial persistence in human hosts. *Nature* **449**, 843–9 (2007).
28. Zheng, N., Whalen, C. C. & Handel, A. Modeling the potential impact of host population survival on the evolution of M. tuberculosis latency. *PLoS One* **9**, e105721 (2014).
29. Blower, S. M. *et al.* The intrinsic transmission dynamics of tuberculosis epidemics. *Nat. Med.* **1**, 815–21 (1995).
30. Porco, T. C. & Blower, S. M. Quantifying the intrinsic transmission dynamics of tuberculosis. *Theor. Popul. Biol.* **54**, 117–32 (1998).
31. Uys, P. *et al.* The Risk of Tuberculosis Reinfection Soon after Cure of a First Disease Episode Is Extremely High in a Hyperendemic Community. *PLoS One* **10**, e0144487 (2015).

32. Menzies, N. A. *et al.* Progression from latent infection to active disease in dynamic tuberculosis transmission models: a systematic review of the validity of modelling assumptions. *Lancet Infect. Dis.* **18**, e228–e238 (2018).
33. Stead W., W. Variation in vulnerability to tuberculosis in America today: random, or legacies of different ancestral epidemics?
34. Barnes, I., Duda, A., Pybus, O. G. & Thomas, M. G. Ancient Urbanization Predicts Genetic Resistance To Tuberculosis. *Evolution (N. Y).* **65**, 842–848 (2011).
35. Raberg, L. & Stjernman, M. The evolutionary ecology of infectious disease virulence. in (eds. Demas, G. E. & Nelson, R. J.) 548–578 (Oxford University Press USA, 2012).
36. Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569–77 (2002).
37. Bergelson, J., Dwyer, G. & Emerson, J. J. Models and data on plant-enemy coevolution. *Annu. Rev. Genet.* **35**, 469–99 (2001).
38. Roy, B. A. & Kirchner, J. W. Evolutionary dynamics of pathogen resistance and tolerance. *Evolution* **54**, 51–63 (2000).
39. Fornoni, J., Núñez-Farfán, J., Valverde, P. L. & Rausher, M. D. Evolution of mixed strategies of plant defense allocation against natural enemies. *Evolution* **58**, 1685–95 (2004).
40. Menzies, D. What Does Tuberculin Reactivity after Bacille Calmette-Guérin Vaccination Tell Us? *Clin. Infect. Dis.* **31**, S71–S74 (2000).
41. Cardona, P.-J. A dynamic reinfection hypothesis of latent tuberculosis infection. Infection **37** (2009).
42. Behr, M. A., Edelstein, P. H. & Ramakrishnan, L. Revisiting the timetable of tuberculosis. BMJ k2738, https://doi.org/10.1136/bmj.k2738 (2018).
43. Stead, W. W. Pathogenesis of a first episode of chronic pulmonary tuberculosis in man: recrudescence of residuals of the primary infection or exogenous reinfection? *Am. Rev. Respir. Dis.* **95**, 729–45 (1967).
44. Cardona, P.-J. Reactivation or reinfection in adult tuberculosis: Is that the question? *Int. J. Mycobacteriology* **5**, 400–407 (2016).
45. Sloot, R., Schim van der Loeff, M. F., Kouw, P. M. & Borgdorff, M. W. Risk of Tuberculosis after Recent Exposure. A 10-Year Follow-up Study of Contacts in Amsterdam. *Am. J. Respir. Crit. Care Med.* **190**, 1044–1052 (2014).
46. Trauer, J. M. *et al.* Risk of Active Tuberculosis in the Five Years Following Infection… 15%? *Chest* **149**, 516–525 (2016).
47. Tiemersma, E. W., van der Werf, M. J., Borgdorff, M. W., Williams, B. G. & Nagelkerke, N. J. D. Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review. *PLoS One* **6**, e17601 (2011).
48. Holmes, C. B., Hausler, H. & Nunn, P. A review of sex differences in the epidemiology of tuberculosis. *Int. J. Tuberc. Lung Dis.* **2**, 96–104 (1998).
49. Horton, K. C., MacPherson, P., Houben, R. M. G. J., White, R. G. & Corbett, E. L. Sex Differences in Tuberculosis Burden and Notifications in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. *PLOS Med* **13**, e1002119 (2016).
50. Neyrolles, O. & Quintana-Murci, L. Sexual Inequality in Tuberculosis. *PLoS Med.* **6**, e1000199 (2009).
51. Ozcaglar, C., Shabbeer, A., Vandenberg, S. L., Yener, B. & Bennett, K. P. Epidemiological models of Mycobacterium tuberculosis complex infections. *Math. Biosci.* **236**, 77–96 (2012).
52. Chamberlain, A. T. Demography in Archaeology, https://doi.org/10.1017/CBO9780511607165 (Cambridge University Press, 2006).
53. Zarulli, V. *et al.* Women live longer than men even during severe famines and epidemics. *Proc. Natl. Acad. Sci. USA* **115**, E832–E840 (2018).
54. Rolff, J. Bateman's principle and immunity. *Proc. R. Soc. London. Ser. B Biol. Sci.* **269**, 867–872 (2002).
55. Zuk, M. The sicker sex. *PLoS Pathog.* **5**, e1000267 (2009).
56. Winterhalter, W. E. & Fedorka, K. M. Sex-specific variation in the emphasis, inducibility and timing of the post-mating immune response in Drosophila melanogaster. *Proc. R. Soc. B Biol. Sci.* **276**, 1109 (2009).
57. Duneau, D. F. *et al.* The Toll pathway underlies host sexual dimorphism in resistance to both Gram- negative and Gram-positive bacteria in mated Drosophila. *BMC Biol.* **15** (2017).
58. Casadevall, A. & Pirofski, L. The damage-response framework of microbial pathogenesis. *Nat. Rev. Microbiol.* **1**, 17–24 (2003).
59. Zumla, A. et al. Towards host-directed therapies for tuberculosis. *Nat Rev Drug Discov*, https://doi.org/10.1038/nrd4696 (2015).
60. Cardona, P. & Cardona, P.-J. Regulatory T Cells in Mycobacterium tuberculosis Infection. *Front. Immunol.* **10**, 2139 (2019).
61. Marzo, E. et al. Damaging role of neutrophilic infiltration in a mouse model of progressive tuberculosis. *Tuberculosis*. **94** (2014).
62. Cardona, P.-J. Revisiting the natural history of tuberculosis: tthe inclusion of constant reinfection, host tolerance, and damage-response frameworks leads to a better understanding of latent infection and its evolution towards active disease. *Arch. Immunol. Ther. Exp. (Warsz).* 58 (2010).
63. WHO. Gender and tuberculosis. (2002).
64. Van den Hof, S., Najlis, C. A., Bloss, E. & Straetemans, M. A systematic review on the role of gender in tuberculosis control (2010).
65. UNDP, U. N. D. P. Discussion paper: Gender and Tuberculosis (2015).
66. Nahid, P. *et al.* Interplay of strain and race/ethnicity in the innate immune response to M. tuberculosis. *PLoS One* **13**, e0195392 (2018).
67. Styblo, K., Meijer, J. & Sutherland, I. Tuberculosis Surveillance Research Unit Report No. 1: the transmission of tubercle bacilli; its trend in a human population. *Bull. Int. Union Tuberc.* **42**, 1–104 (1969).
68. Coscolla, M. *et al.* M. tuberculosis T Cell Epitope Analysis Reveals Paucity of Antigenic Variation and Identifies Rare Variable TB Antigens. *Cell Host Microbe* **18**, 538–48 (2015).
69. Heimbeck, J. Incidence of tuberculosis in young adult women with special reference to employment. *Br. J. Tuberc.* **32**, 154–166 (1938).
70. Bloom, B. R. New Promise for Vaccines against Tuberculosis. *N. Engl. J. Med.* **379**, 1672–1674 (2018).
71. Boldin, B., Geritz, S. A. H. & Kisdi, É. Evolutionary ecology research. Evolutionary Ecology Research 11, (Evolutionary Ecology, 1999).
72. Trauer, J. M. *et al.* Risk of Active Tuberculosis in the Five Years Following Infection…15%? *Chest* **149**, 516–525 (2016).
73. Kaplan, H., Hill, K., Lancaster, J. & Hurtado, A. M. A theory of human life history evolution: Diet, intelligence, and longevity. Evol. Anthropol. *Issues, News, Rev.* **9**, 156–185 (2000).
74. Galor, O. & Moav, O. The Neolithic Origins of Contemporary Variations in Life Expectancy. *SSRN Electron. J.*, https://doi.org/10.2139/ssrn.1012650 (2007).
75. Marino, S., Hogue, I. B., Ray, C. J. & Kirschner, D. E. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* **254**, 178–96 (2008).

Acknowledgements

The project leading to these results has received funding from "la Caixa" Foundation (ID 100010434), under agreement LCF/PR/GN16/10290002.

Author contributions

P.J.C. has conceived the work, wrote and structured the paper, and designed the TBOREX model and a basic coinfection model; C.P. has refined and implemented the TBOREX model in Matlab; M.C. has designed the final version and implemented in Matlab the coinfection model together with C.P. and run the sensitivity analysis; All authors have worked in the analysis and interpretation of the data.

Competing interests

PJC is founder of the companies Archivel Farma and Manremyc, devoted to the prophylaxis and treatment of tuberculosis. CP and MC declare no potential conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-56769-1.

Correspondence and requests for materials should be addressed to P.-J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

9 Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers

9.1 Summary

The main goal of this article is to estimate the real incidence that the so-called first COVID-19 wave (March to May 2020) had in the main European states. During this so-called first wave, governments and health ministries were not prepared for the storm that was coming and they reacted to the health system collapse by forcing their citizens not to move from their home. The information systems were not prepared, and there was a lack of testing capacity for the detection and monitoring of COVID-19 cases. In fact, only the severe cases that went to the hospital were detected. Estimating real incidence was important to know which percentage of the population had antibodies, as this fraction of population would be less susceptible to having COVID-19 again or, at least, a severe disease if they came into contact with the virus again.

One of the first points discussed in the article is determining the infection fatality rate (IFR), i.e., how many people of those that have COVID-19 are going to die due to the infection. This is different from case fatality rate (CFR), which is computed from the detected individuals. From available data (at the time), it was determined that it was between 0.5 and 2%; in this paper our central hypothesis is that the COVID-19 IFR is 1%.

Secondly, we discuss the possible deviations between different countries. On the one hand the possible variability of IFR in different European countries is discussed. It is known that the main COVID-19 fatality factor is age, so countries with different population pyramids are going to have different IFRs. In general, it is seen that, although European population pyramids are slightly different, we may suppose that all the countries have the same IFR. However, another possible bias is the reporting deaths criteria. For example, this is a problem with Belgium, which reports suspected COVID-19 deaths, while other countries report deaths with a COVID-19 positive test. It has been seen that the excess mortality is higher than that reported for COVID-19 [157]. Further analysis can be done when all excess mortality is measured, if we are able to distinguish which is the part due to COVID-19.

There is also a delay between cases and deaths that is different for each country [64]. We suppose that different delays are due to a different diagnosis delay in each country. From a correlation analysis we identify this diagnosis delay in particular for each country.

As for estimating diagnosis percentage, we can use the last available detection percentage to estimate the real pandemic indicators for the different countries. In this sense, we may find some countries that have similar reported incidences but different estimated incidences, and the other way around, i.e., countries with different reported incidences that have a similar estimated incidence.

The results of this paper were confirmed by a Spanish seroprevalence study (ENE-COVID) [138]. In fact, from this study we can conclude that for Spain the COVID-19 IFR is 0.83%, very similar to the 1% central hypothesis made in our article.

9.2 Main contributions

The main contributions of this article are:

- COVID-19 infection fatality rate is estimated from bibliography review to be around 1% IFR in European countries.
- We decided upon criteria to chose the European countries used to perform the analysis. Possible bias between these different countries is analysed. There are no great differences in the population pyramid and criteria for reporting COVID-19 deaths are the same in all countries except Belgium. A possible bias is Belgium is estimated at 50%.
- Total number of COVID-19 cases in each country during the first wave is estimated. For each country, we estimated the delay between detection and death, then the delay between case and detection was deduced.
- From delays and IFR estimations, we computed the diagnosis percentage of the first wave and its evolution. The last estimated diagnose percentage is used to estimate active cases in each country and compute different epidemiological indicators with the estimated active cases.
- A new epidemiological indicator is presented: effective potential growth (EPG) as the dot product between the active case incidence

(A14) and the empirical reproduction number (ρ_7). It is a short time prediction of A14 in the following days.

9.3 Application and further work

The methodology explained in this article was used to determine Spanish COVID-19 incidence before seroprevalence study results were released [168, 51, 145, 59]. It has been used during the successive reports made by Computational Biology and Complex Systems groups (BIOCOM-SC) of the Universitat Politècnica de Catalunya (UPC) of which I am member [43, 37, 42]. In fact, a draft version of the methodology explained in this paper has been used since report 1 delivered on 15 March 2020 to the DG-Connect European Commission department [38].

At this moment, detection and contact-tracing in Europe has improved and detection rate is high [37], so this model is not as useful as it was in the first wave. Also, the vaccination process is effective and IFR is reduced because of it [25]. This can be taken into account to modify the model. There are many countries where detection and contact-tracing are not well implemented; in these countries our model is still very useful to estimate the number of infected people.

This model has been improved taking into account the deaths by age range and IFR for each age range. Then, differences in population pyramids do not affect the model because, in fact, each country has its own IFR. Nowadays, vaccination protection and vaccines administered by age group have also to been taken into account. Effective potential growth index was implemented and used in several reports delivered to the Catalan health authorities [45, 2]. EPG is used in Catalonia as "índex de rebrot", one of the main indexes to study epidemic stage.

9.4 Article

The full article can be read online at:
Català M, Pino D, Marchena M, Palacios P, Urdiales T, Cardona PJ, Alonso S, López-Codina D, Prats C, Alvarez-Lacalle E. Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. *PLOS One.* 2021 Jan 7;16(1):e0243701. `https://doi.org/10.1371/journal.pone.0243701`

Supplementary information can be found at: `https://doi.org/10.1371/journal.pone.0243701.s001`

PLOS ONE

RESEARCH ARTICLE

Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers

Martí Català[1,2☯], David Pino[1☯], Miquel Marchena[1], Pablo Palacios[1], Tomás Urdiales[1], Pere-Joan Cardona[2,3,4], Sergio Alonso[1], David López-Codina[1], Clara Prats[1,2], Enrique Alvarez-Lacalle[1☯*]

1 Department of Physics, Universitat Politècnica de Catalunya (UPC · BarcelonaTech), Barcelona, Spain, **2** Comparative Medicine and Bioimage Centre of Catalonia (CMCiB), Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol (IGTP), Badalona, Catalonia, Spain, **3** Experimental Tuberculosis Unit (UTE), Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol (IGTP), Universitat Autònoma de Barcelona (UAB), Badalona, Catalonia, Spain, **4** Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain

☯ These authors contributed equally to this work.
* enric.alvarez@upc.edu

OPEN ACCESS

Citation: Català M, Pino D, Marchena M, Palacios P, Urdiales T, Cardona P-J, et al. (2021) Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. PLoS ONE 16(1): e0243701. https://doi.org/10.1371/journal.pone.0243701

Editor: Joan A Caylà, Agencia de Salut Publica de Barcelona, SPAIN

Received: April 23, 2020

Accepted: November 27, 2020

Published: January 7, 2021

Data Availability Statement: Data is available via ECDC repositories on COVID-19: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

Funding: CP, PJC and MC received funding from La Caixa Foundation (ID 100010434), under agreement LCF/PR/GN17/50300003; PJC received funding from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Grup Unitat de Tuberculosi Experimental, 2017-SGR-500; CP, DL,

Abstract

Policymakers need clear, fast assessment of the real spread of the COVID-19 epidemic in each of their respective countries. Standard measures of the situation provided by the governments include reported positive cases and total deaths. While total deaths indicate immediately that countries like Italy and Spain had the worst situation as of mid-April, 2020, reported cases alone do not provide a complete picture of the situation. Different countries diagnose differently and present very distinctive reported case fatality ratios. Similar levels of reported incidence and mortality might hide a very different underlying pictures. Here we present a straightforward and robust estimation of the diagnostic rate in each European country. From that estimation we obtain a uniform, unbiased incidence of the epidemic. The method to obtain the diagnostic rate is transparent and empirical. The key assumption of the method is that the infection fatality ratio of COVID-19 in Europe is not strongly country-dependent. We show that this number is not expected to be biased due to demography nor to the way total deaths are reported. The estimation protocol is dynamic, and it has been yielding converging numbers for diagnostic rates in all European countries as from mid-April, 2020. Using this diagnostic rate, policy makers can obtain Effective Potential Growth updated every day, providing an unbiased assessment of the countries at greater risk of experiencing an uncontrolled situation. The method developed has been and will be used to track possible improvements in the diagnostic rate in European countries as the epidemic evolves.

Introduction

The evolution of the epidemic in Europe has affected Spain and Italy more strongly than other countries so far. This is clear from reported cases and fatalities in these countries [1–3].

SA, MC received funding from Ministerio de Ciencia, Innovación y Universidades and FEDER, with the project PGC2018-095456-B-I00. EA-L received funding from Spanish Ministerio de Economía, Industria y Competitividad under grant number SAF2017-88019-C3-2-R. This project has been partially funded by the European Comission - DG Communications Networks, Content and Technology through the contract LC-01485746

Competing interests: The authors have declared that no competing interests exist.

However, comparative assessment of the spread of the pandemic in other European countries has been more difficult to make. The reason is that the real incidence of the epidemic in each country cannot be known with certainty, because countries are not able to perform the same number of polymerase chain reaction tests (PCR) and consequently the comparison of the ratio of those infected is difficult [4]. Policy responses have also differed, with some countries focusing on clinical testing in hospitals, while others have tried to use tests, at least partially, to determine some local chains of transmissions [5, 6]. The lack of clear inter-country comparison in Europe has deep implications for the future structure of the European Union since many decisions are taken based upon the sense of gravity in the country in question. For these reasons, it is important to have a proper measure of the relative spread of the epidemic. Policymakers must know what the real situation in their own countries is in comparison to others so that their decisions on the future of reopening and economic reconstruction are taken based not on false impressions, but on data. In this sense, policymakers must perceive the method as unbiased, simple, and robust. Most importantly, the relative comparisons between countries must be as shielded as possible from the hypothesis of the method. In this sense, methods have recently been developed [7–9] in order to assess the situation inferred from data. This work has been instrumental in providing a better picture of the situation. However, the work lacks the recipe-type nature needed to direct a policy response.

The focus of this paper is, therefore, to introduce a method to compute the real diagnostic rate and the real incidence of COVID-19 in each European country, testing whether the key hypothesis of the method is fulfilled and, if slightly off, whether it would affect all countries in the same way. In other words, we provide a recipe for policymakers that we have shown to be correct, and unbiased across countries and useful to make inter-country comparison, provided the evolution and prognosis of the disease in a patient is not strongly dependent on socio-economic factors, and only on age, sex and previous clinical history.

We must recall here that the ability to determine the diagnostic ratio is essential to evaluate what the real number of infected people is. Knowledge of this number is not only useful to visualize the full scope of the epidemic but also to properly estimate the number of people with probable short-term immunity. In this sense, our method can be added as an empirical take on other assessments of the real incidence of the disease and to study the possibility of developing herd immunity. A large number of real infected people would be a positive scenario for policymakers while a low number would be negative. It is thus very important to err on the side of caution in all our estimates, settling on the less optimistic take.

The basic structure of the paper is the following. First, we give a general overview of our framework in the methods section. Then we discuss our key assumption: the infection fatality ratio (IFR) in European countries experiencing a significant incidence will be roughly the same, given the similar structure of the population. If IFR were to be lower, or higher, it would affect all countries in the same way and would not affect most policy decision-making since it would move all countries in the same direction. We take this IFR to be 1% and proceed to test whether, effectively, there is a strong correlation between the day of reported deaths with the number of cases taken 7-10 days before. Once a given value for the IFR is taken, one must consider that people do not die immediately from the disease, as it takes roughly 18 days after infection [10–12]. In other words, the present values of the death toll can provide an estimation of the number of infected people 18 earlier. Knowing the number of infected people at present, and not 18 days in the past, is crucial. We attack this problem considering that people who become infected are usually diagnosed a few days after the onset of the symptoms, which can be 8 to 14 days after infection occurs. By comparing the number of people diagnosed on a certain date with our estimation of the real number of infected people, we can estimate what

percentage of the cases is being diagnosed. We can then calculate this for different countries and regions and test how this ratio has changed dynamically as the epidemic advanced.

In the results section, we provide a full detailed description of how this fraction has become steady in the last weeks. We demonstrate that the percentage of diagnosis throughout the development of the epidemic has taken values that gradually converge for most countries. This gives a final clear picture showing the rate of diagnosis for each country. Using this rate it is straightforward to provide a present-day estimate of the incidence given the number of reported infected people in each country as long as we can observe that the rate of diagnosis remains fairly constant. For policymakers, we have constructed an index called Effective Potential Growth (EPG) that combines this information with the growth rate of the epidemic to provide insight regarding which countries are, comparatively and in the short-term, in the potentially most complicated situation [9]. Finally, we analyze the sensitivity of the EPG index to variations of the different parameters not based in previous studies and we show how robust is the index in predicting and increase of the incidence during the first phase of the pandemic and in detecting secondary outbreaks.

Methods

Framework of our methodology

Our analysis will be applied to European countries with a minimum of 500 deaths on April 15, 2020 so that we can guarantee a minimum statistical significance, so that the fluctuations $1/\sqrt{N}$ of an associated binomial distribution are below 5%. The analyzed countries are: Belgium, France, Germany, Italy, Netherlands, Portugal, Spain, Sweden, Switzerland and the United Kingdom. Our two core assumptions are that the IFR in all European countries is roughly the same and that reported data of death due to COVID-19 is uniform in all European countries under consideration. We will address these two hypotheses in the following sections. With these assumptions we need to carry out four steps, as indicated in Fig 1, to obtain the percentage of diagnosis. First, using a common reference IFR = 1% and, given the reported death count, we estimate the number of cases 18 days ago. According to medical reports people die between 15 and 22 days after the development of the first symptoms [13]. This time to death, TtD = 18 days, after the development of the first symptoms will not be country-specific for demographic reasons. The estimated number of infected people with the disease at time t, E_t, (see process in Fig 1(A)) reads:

$$E_t = \frac{d_{t+\mathrm{TtD}}}{\mathrm{IFR}}, \tag{1}$$

where $d_{t+\mathrm{TtD}}$ is the number of reported deaths at time t + TtD.

This allows us to estimate the number of cases TtD = 18 days ago. This value can be compared with the number of cases detected 18 days ago, obtaining a diagnostic percentage. This result is an unrealistic lower bound because no one performs PCR testing the first day of symptoms; this is usually done much later. Actually, people normally do not call a doctor at the first symptoms. Furthermore, depending on the availability of tests, saturation of the health system, and other external factors, countries show great variability in the time of diagnosis delay.

Countries accumulate some delay that may reach 18 days, in the case of a country detecting people as late as death. This delay to detection (DD) due to lags in diagnosis corresponds to the time between the patient having the first symptoms and being reported by the health system. In fact, this time in some countries may vary throughout the course of the infection. Therefore we cannot assume that the estimated and the reported data are comparable and we need to know what the diagnostic time was for each of the countries studied.

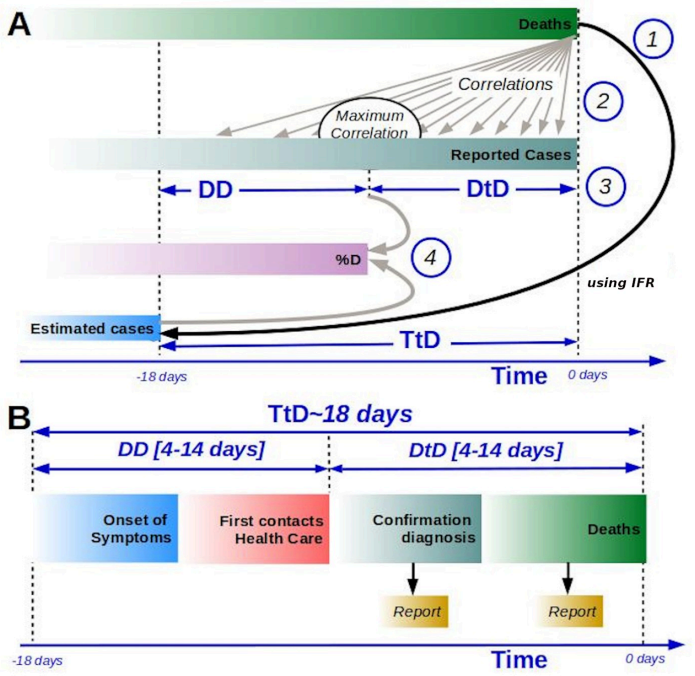

Fig 1. General framework of the calculation of the percentage of diagnosis. (A) Processes involved in the calculation of the percentage of diagnosis: 1. Evaluation of the estimated cases using TtD and IFR, 2. Calculation of time correlation between reported cases, C_t, and reported deaths, d_t, 3. Evaluation of the time between diagnosis and death (Diagnosis to Death DtD) by the maximum of correlation (country dependent), with DD as the Delay to Detection, and 4. Evaluation of percentage of diagnosis based on estimated cases and reported cases. (B) Standard evolution of casualties by COVID-19, from the onset of the symptoms to death; times to reported cases and deaths are shown. Time-lines in the figure are not proportional to real-time delays.

https://doi.org/10.1371/journal.pone.0243701.g001

We can compare the reported deaths with the reported cases to find the maximal correlation; see process 2 in Fig 1(A); to estimate the DD, see process 3 in Fig 1(A). Finally the ratio between reported cases at DD and estimated cases, see below, provides an estimation of the percentage of diagnosis; see process 4 in Fig 1(A). Note that the usual development of the reporting of a new case/death, see Fig 1(B), depends on the particular country under consideration, which determines DD. In fact, DD also includes a delay in reporting, from the diagnosis to death, to official information systems.

IFR of COVID-19 in Europe

The cornerstone of our analysis is that the IFR in all European countries will not be biased against any country in particular. We should point out immediately that we are not arguing that there are not important uncertainties in the IFR. What we do claim and check in this methodology is that these uncertainties will not generate any biased against particular countries and should not affect policy decision. We take the IFR of COVID-19 in Europe to be between 0.3-3% and we assume 1% to be the benchmark scenario.

This value (1%) is the case fatality ratio (CFR) observed in the initial stages of the pandemic in China, Republic of Korea (South Korea) and the Diamond Princess cruise. In these three locations, it was found to be around 1-2.6%, and error margins came from different sources, respectively [14–16]. In China, once adjusted for demography and under-ascertainment, case fatality ratio during February, 2020 was estimated to be 1.4%. In South Korea, the ability to test all the population in contact with infected people and the tracking of contagious chains was thorough. Despite this, the reported CFR increased from initial values around 0.5-0.7% to higher values around 2%. In the Diamond Princess cruise, CFR for confirmed cases was 2% but estimation of false negatives and the possibility that a fraction of the passengers never developed symptoms and was never tested put the CFR again around 1%. South Korea and the Diamond Princess cruise provide complementary evidence, one coming from a natural experiment and another from a country with the ability to perform half a million tests/day from the very beginning of the transmission chain [17]. If we accept the two measurements of the CFR independently, the most likely interval of IFR is between 0.5 and 2%.

Experimental results from random testing in the German city of Gangelt [18] and preliminary results from Iceland [19, 20] indicate the presence of a layer of people fully asymptomatic people that are normally not detected. This group of people have passed the disease without any knowledge seems to be larger than previously thought. These preliminary studies point to a CFR of around 0.5% in zones where the epidemic was not fully spread. We cannot disregard the possibility that, just as CFR increased with time even in South Korea, similar studies in countries with more cases, could reveal a higher CFR.

Recently, other authors, using different techniques, have obtained estimates for CFR or IFR in different countries/regions. [21] obtained the cumulative incidence of SARS-CoV-2 infection in New York State based on dry-blood spot SARS-CoV-2 antibody reactivity performed on 15,000 individuals over 18 years from March 19, 2020 to 28 March, 2020. They found that the estimated cumulative incidence was 14% (2.7 million people), which gives a CFR of 0.5%. [22] uses a mathematical model with a cohort analysis approach to determine the range of case fatality ratios in Hubei province (China) from January 22, 2020 to March 11, 2020. They demonstrated that CFR is from 4.8% to 6.1%. [23] statistically estimate the incidence of the pandemic in 70 countries by using indirect reporting, where the questions a participant answers are not about herself (*network scale-up method*). They found Brazil, Ecuador and Ukraine CFR are 1% on May 17, 1.61% on April 15 and 0.56% on April 26, 2020, respectively. Finally, [24] analyze data from 139 countries to conclude that the global IFR is 1.04%.

It is thus reasonable to consider IFR at 1% as an easy policy guiding principle and not to use the more positive scenario of 0.5%.

Unbiased nature of IFR in Europe

There are three sources of possible biased IFR across countries. The disease affects older people with comorbidity problems more than to healthy younger ones, and men diagnosed with COVID-19 are more likely to die than women [25], the same way it happened in previous coronavirus epidemics [26, 27]. Across all European countries, the male/female ratio is approximately the same except for the oldest people. For people above 80 years, sex ratio (number of men by 100 women) in Europe oscillates between 85 (Albania) and 33 (Latvia) [28]. This is precisely the group with the highest case fatality ratio. It is thus very important to assess how the different demographic structures of European countries could affect our central benchmark [29]. The same must be said about the relative prevalence of other comorbidity factors. We proceed to show that, with the data we have today, and the demographic and comorbidity structures, none of these possible sources of bias can have anything but a small effect. To

Table 1. For each age group: Number of cases and deaths reported in South Korea on April 15, 2020; percentage of population, COVID-19 cases and deaths, and case fatality ratio.

Age group	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	≥80
Cases	136	576	2909	1135	1411	1942	1342	704	480
Deaths	0	0	0	1	3	14	33	68	111
Population (%)	8.1	9.3	13.1	13.8	16.0	16.5	12.6	6.9	3.6
Cases (%)	1.3	5.4	27.4	10.7	13.3	18.3	12.6	6.6	4.5
Deaths (%)	0	0	0	0.4	1.3	6.1	14.4	29.6	48.3
CFR (%)	0	0	0	0.09	0.21	0.72	2.46	9.66	23.13

Source: United Nations World Population Prospects 2019 and Korea Centers for Disease Control and Prevention.

https://doi.org/10.1371/journal.pone.0243701.t001

do so, we will use and make a comparison with the CFR of South Korea on April 15, 2020, of 2.16%.

Table 1 shows, for each analyzed age group, the demographic structure of South Korea, the number of cases and deaths, the corresponding percentage, and the corresponding CFR officially reported on April 15, 2020. As may be observed in the table, by comparing the percentage of population and cases for all the age groups we can conclude than in South Korea people below 49 years old are infected/detected less than their corresponding population importance. The contrary occurs for the people above 50 years old. People over 80 present an increase in infection of 32% with respect to their population importance. This is probably because they present increased symptoms and are therefore tested more often. However, recent studies [30, 31] show that younger people are less susceptible to be infected. Regarding deaths, these differences are even more important, and the change occurs for the age group 60-69.

To analyze what the role played by the differences in demography in Europe in the COVID-19 cases and fatalities is, we have obtained from the United Nations World Population Prospects 2019 the demographic distribution by age in the considered countries (see Table 2).

We can readily assess that, compared with South Korea, all the countries have a larger percentage of population over 70 years. The percentage is even larger if people over 80 is considered (108% larger for Italy). It is also important to note that the relative differences in each of the cohorts between the European countries shown in the table is small. Only Italy presents a relevant larger than average ratio of people over 80.

To analyze the role played by demography, we use this demographic data and the number of cases and deaths reported by South Korea on April 15, 2020 (see Table 1) to estimate the

Table 2. Percentage of population by age group for the analyzed European countries.

Country/Age group	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	≥80
Belgium	11.2	11.3	12.0	13.0	13.1	13.8	11.8	8.1	5.7
France	11.5	12.1	11.3	12.3	12.8	13.2	11.9	8.8	6.2
Germany	9.4	9.5	11.2	13.0	12.2	16.1	12.7	8.9	7.0
Italy	8.3	9.5	10.1	11.6	14.9	15.8	12.4	10.0	7.5
Netherlands	10.2	11.4	12.2	12.2	12.6	14.7	12.4	9.3	4.9
Portugal	8.2	10.0	10.5	11.9	15.5	14.5	12.7	10.0	6.7
Spain	9.1	10.1	9.9	12.6	17.0	15.1	11.4	8.6	6.3
Sweden	11.8	11.2	12.6	13.1	12.5	12.8	10.8	9.8	5.3
Switzerland	10.2	9.6	12.0	14.1	13.5	15.3	11.3	8.7	5.3
United Kingdom	11.8	11.3	12.6	13.7	12.7	13.5	10.7	8.6	5.1

Source: United Nations World Population Prospects 2019.

https://doi.org/10.1371/journal.pone.0243701.t002

Table 3. Estimated relative CFR assuming these countries have the same CFR by age group as reported by South Korea on April 15, 2020 (see Table 1), and officially reported relative CFR on that date. CFR of South Korea on April 15, 2020 was 2.16%. The officially reported CFR on that date for each country is indicated in parentheses. Source: European Center for Disease Prevention and Control.

Country/Relative CFR	Estimated from South Korea reported data on April 15, 2020	Reported on April 15, 2020
Belgium	1.35	6.18 (13.4)
France	1.47	7.02 (15.2)
Germany	1.56	1.18 (2.6)
Italy	1.67	6.00 (13.0)
Netherlands	1.29	4.97 (10.7)
Portugal	1.56	1.50 (3.3)
Spain	1.46	4.63 (10.0)
Sweden	1.35	4.17 (9.0)
Switzerland	1.31	1.62 (3.5)
United Kingdom	1.28	6.93 (15.0)

https://doi.org/10.1371/journal.pone.0243701.t003

CFR corrected by population for each European country. To obtain it, first, we estimate the number of cases for each European country E and age group A as follows:

$$\text{Estimated Cases in } E_A = \text{Cases South Korea}_A \times \frac{\%\text{Population } E_A}{\%\text{Population South Korea}_A},$$

and the same for the corresponding number of deaths. That is, we are assuming that the CFR for each age group of all the European countries is the same as shown in Table 1 for South Korea. However, this doesn't imply necessarily that the country CFR for a European country is the one reported by South Korea because the number of cases and deaths for each age group and country is different in each country. Table 3 shows the CFR obtained by performing this analysis and the CFR officially reported at the different European countries on the same date. Both values are presented relative to the CFR reported by South Korea on April 15, 2020, of 2.16%.

As may be observed in the first column, when we only consider demographic differences between countries, the difference between the worst and best case of the relative CFR is around 30%. Most countries are in the range between 1.28% and 1.5%, with the average of the relative CFR being 1.34%. Therefore, CFR for all the countries except Italy is, at most 20% from the average value and typically around 10%.

Analysis of possible bias in reported deaths due to COVID-19

An unrelated source of bias in the estimation of the real COVID-19 cases, is the possibility that different countries differ in how they treat and count the population that dies after having a very bad prognosis. We know this group is strongly affected by the virus [32]. In blunt terms, we must examine the possibility that different countries are counting the raw number of deaths differently.

Before entering in the details of the analysis, let us point out that two indications go against this possibility. First, health care systems in Europe have different resources in different countries with differing focuses and priorities, but they attend anyone with COVID-19 with the exception of possible patients with multifactorial problems who might be in very fragile conditions. Elderly people in nursing homes who die under suspicious situations are uniformly not reported following European Center for Disease Prevention and Control (ECDC) advice. There is a single exception that we know of: Belgium [33]. Belgium seems to be reporting

unconfirmed cases from nursing homes without tests as due to COVID-19. It is quite clear that this includes a good number of people who, either, did not die from COVID-19 or for whom COVID-19 was not an important factor in the prognosis. Therefore, we will include a reminder that Belgian data are biased compared with other countries, being 48% lower on April, 20 2020 [34], given the number of reported deaths from nursing homes compared with hospitals. There is a second argument regarding the treatment of the elderly population in other countries. If large undercounting is the case, it should be noted in the case fatality ratio for people 80 years and older, which is not observed in the countries from which we have data.

Regarding the possible COVID-19 mortality not reported in the analysis due to a lack of PCR tests, we analyze here the situation in Spain. The National Epidemiology Center (Instituto de Salud Carlos III) of Spain published the results of the Daily Mortality Monitoring System (MoMo) for June 17, 2020 [35]. They evaluated which periods had mortality well above the average for previous years. When evaluating the period from March 13 to May 22, 2020 for the whole of Spain, they saw that, as expected, mortality was much higher than in previous years. An increase of 55.8% was observed. However, it is interesting to compare this with the data reported for COVID-19 deaths. The reported deaths by COVID-19 were roughly 28,000 and the reported excess of deaths by the MoMo surveillance system was 44,000. We think that the assessment of around 35% underreporting can be taken indeed as a worst-case scenario for a highly impacted country. It seems reasonable to expect other countries to have underreported at around this level 15-35% [36, 37]. All the data suggest right now, that the undercounting due to a different treatment of the very fragile population is highly unlikely across Europe, and at most introduces changes in IFR of ±10%.

Treatment of inter-country bias in diagnostic time-delays

Having shown that the IFR should not represent a bias in European countries larger than 25%, we now address the question of how to deal with the real sources of bias in the diagnostic rate for each country. To estimate DD we look for a correlation between the number of reported cases (see Fig 2A) and the number of reported deaths (see Fig 2B) [1, 38]. To deal with noise effects we put a weighted moving average filter on the data both of cases and deaths. The correlation time between reported cases and reported deaths will be called time from diagnosis to death (DtD), and:

$$\mathrm{TtD} = \mathrm{DD} + \mathrm{DtD}. \tag{2}$$

In Fig 2C we see the correlation [39] between reported cases and reported deaths assuming different DtD for Germany, Spain, and Switzerland. As you might expect, correlations have values close to 1. In most cases the correlation has a concave parabolic shape with a clearly defined maximum. We assume this maximum represents DtD for each country. The correlation interval is estimated as the points where the correlation is greater than 99% of the observed maximum. We decided to set a lower limit of 4 days and an upper limit of 14 days [11] because we believe that time outside this range would be unrealistic. For countries that have not seen a clear correlation (Sweden) it was decided to explore the entire DtD interval (4 to 14 days). Approximate values for DtD are shown in Fig 2D. In the S1 File the correlation curves can be seen for the 10 analyzed countries (see S1 Fig in S1 File). Then, by using Eqs (1)

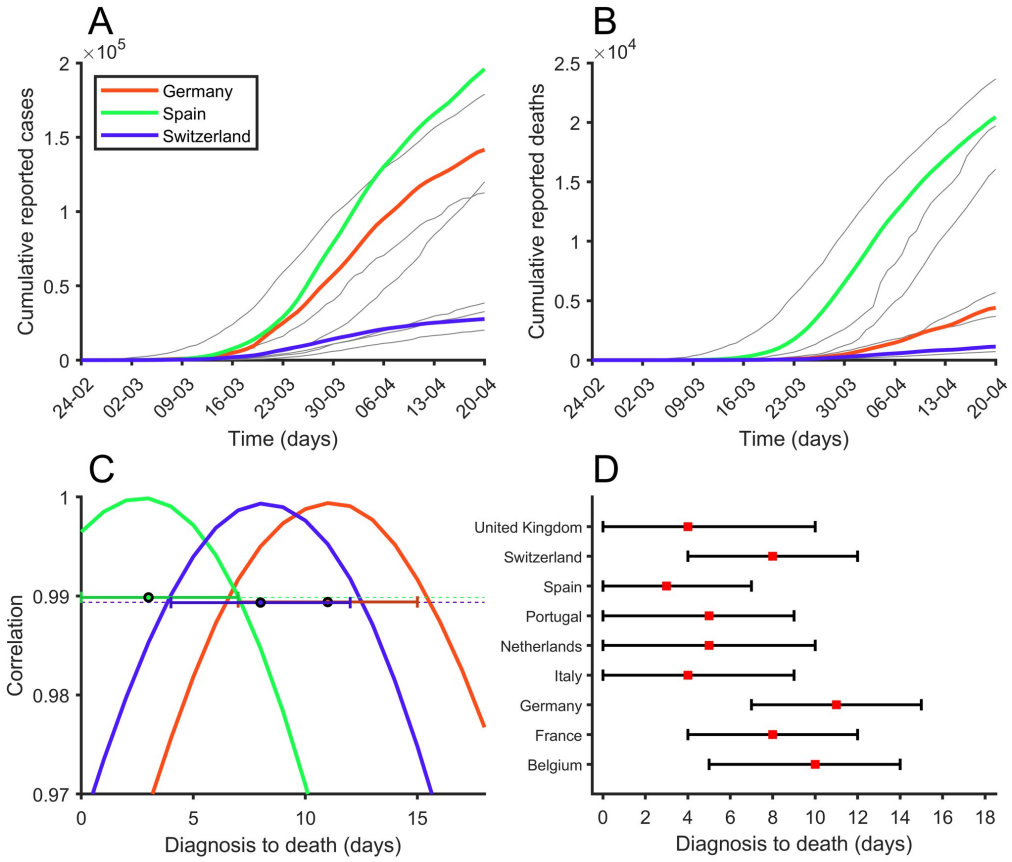

Fig 2. Correlation between reported cases and deaths. (A) Number of cumulative reported cases, (B) Number of cumulative reported deaths and (C) Correlation between reported cumulative cases and reported cumulative deaths exploring different delays between diagnosis (reported) and death, for Germany (red), Spain (green), and Switzerland (blue). (D) Maximum correlation is marked with a red square for each country. 99% correlation interval can be seen with black bars.

https://doi.org/10.1371/journal.pone.0243701.g002

and (2), the percentage of diagnosis cases diagnosed at time t, $\%D_t$, reads:

$$\%D_t = \frac{E_{t-\mathrm{DD}}}{C_t} = \frac{d_{t+\mathrm{TtD}-\mathrm{DD}}}{\mathrm{IFR}\ C_t} = \frac{d_{t+\mathrm{DtD}}}{\mathrm{IFR}\ C_t}, \tag{3}$$

where C_t is the number of reported cases at time t.

Results

Diagnostic rate by country

As discussed in Methods, we use the same IFR = 1% in all European countries instead of making small corrections for demography. The bias due to demography was shown to be around 10%-15%, precisely the same order of magnitude we obtained for the possible bias in the

counting of reported mortal cases. Given that our aim was to provide a clear method for policymakers and that there are no data on how, or even if, the two correlate, a common IFR allows us to homogenize the results with the clear limitation that we will obtain reasonable estimations and not exact results. The resulting picture is expected to be closer to reality than that obtained using purely reported data, but worse than that correcting properly for age and diagnosis if the data of IFR for all age brackets and locations (nursing homes, hospitals, individual homes) were available, which is not the case.

The estimation of the diagnostic rate is straightforward. From the cumulative number of deceased each day, and multiplying by 100 (1% IFR), we get the cumulative number of people with symptoms 18 days ago [10–12] simply by rescaling and displacing the cumulative death curve of any country backwards in time. To give an initial realistic and homogenous diagnostic rate we must establish how many days are needed as a bare minimum to be able to detect a patient from the onset of symptoms. First, the patient has to feel sufficiently sick and then contact the health service. From this contact, the doctor needs to be suspicious that the person has the disease and request a test. Then, this test must be available, performed, and the result received and noted. It is clear that a bare minimum of one week is needed for this process. We use the term 7-Days Diagnostic Rate (7D-DR) for the diagnostic rate with a benchmark of one week of diagnosis delay.

We explain the procedure to get 7D-DR for each country in Fig 3A. We take the cumulative death curve of any country, rescale it according to the IFR, and displace it towards the past 11 days. This is 18 days back to the onset of symptoms and then 7 days forward to be detectable/diagnosable. From this curve, we can obtain the rate between the cumulative number of people who had symptoms for 7 or more days and the cumulated number of people detected 11 days ago. It is thus clear that this homogenous analysis across countries could be performed assuming 5D-DR or 9D-DR and different IFRs. It yields a proper first estimation of the situation.

We argue, however, that there is indeed bias in the way people deal with the health care system in normal situations, and especially under epidemic circumstances. Different countries and populations in fact behave very differently. We have observed that this is the case in the Methods section checking the delay between diagnosis and death using time-displaced correlation analysis. This is the reason why we also define the Delay-to-Detection Diagnostic Rate (DD-DR) as the diagnostic rate computed using a different time delay between the appearance of symptoms and detectability for each country. We proceed to use Fig 4, with Spain as an example, to explain the concept behind DD-DR.

For Spain, the maximum correlation between cumulative death curves and cumulative reported cases appears when cumulative deaths are displaced 4 days backwards. This suggests a DD of around two weeks (18 − 4 = 14 days). This makes sense in a situation like that in Spain during March, 2020. The population receiving news that the health care system is under stress may decide to delay reporting of symptoms unless these are very serious. Additionally, there is the possibility that tests are not available to people who report to primary health care centers with symptoms, and that the delay between the test, its positive result, and its recording in official information systems is not negligible as well.

It is thus important to correct for this bias in the estimation of the diagnostic rate. It is clearly not the same to have a time delay from symptom to the detection of 14 days and 7 days. DD-DR can be computed from Spain just as we did before for the 7D-DR using the same rescaling of the cumulative death curve as before but using a displacement backwards of 4 days instead of 11 days. Fig 4 shows how the DD-DR is obtained in different countries depending on the delay between symptoms and detectability. Countries with a lower DD, such as Germany, have the same 7D-DR as DD precisely because they diagnose as early as realistically possible.

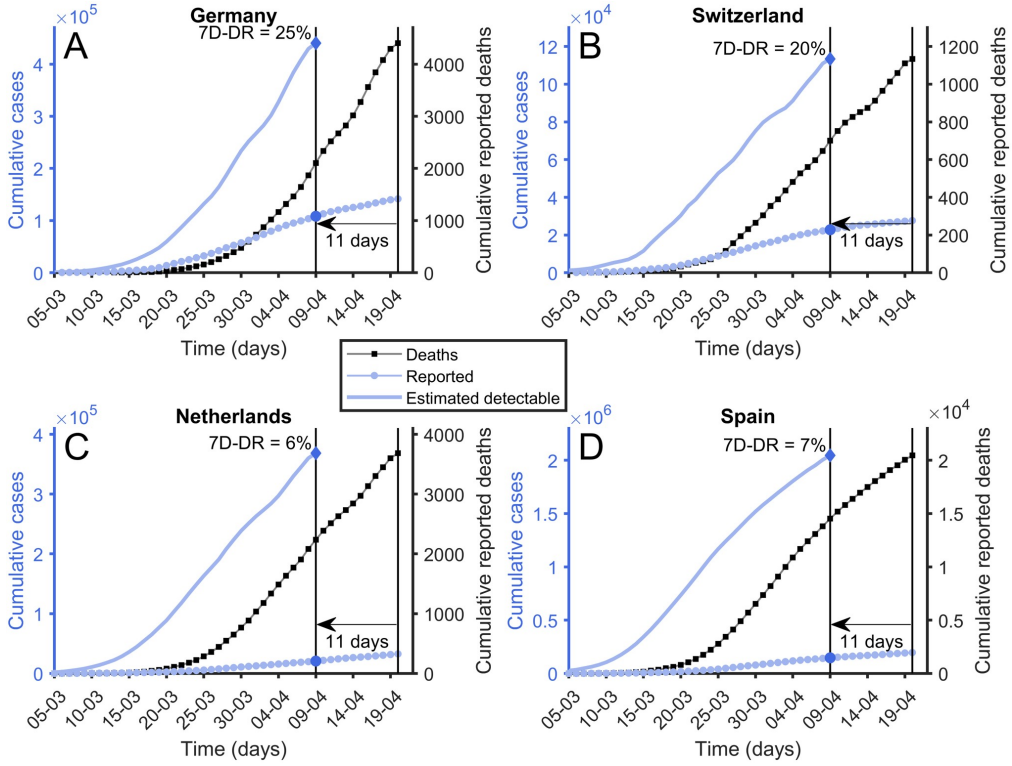

Fig 3. 7-Days diagnostic rate. Reported cumulative number of deaths (black squares), reported cumulative number of cases (blue circles) and estimated number of cases calculated using Eq 1 (solid blue line). To compute 7-Days Diagnostic Rate a diagnosis-to-death time of 11 days is used. Its value is calculated using the latest available points. (A) Germany. (B) Switzerland. (C) Netherlands. (D) Spain.

https://doi.org/10.1371/journal.pone.0243701.g003

We note now that both 7D-DR and DD-DR can be tracked over time; as the epidemic advances we can check how these diagnostic rates changes. Each new day we can look 11 days back for the 7D-DR and compute the diagnostic rate. DD-DR can be tracked similarly. In Fig 5 we show the evolution for both as a function of time for three selected countries. We observe that the DD-DR reaches a steady state after the initial stages of the disease while 7D-DR seems to be more affected by trends. This is to be expected since DD-DR uses, precisely, the maximum correlation delay and so it fluctuates less. The DD-DR is not only more stable but it also allows a proper assessment of the errors involved. The main error is the fact that the exact delay between onset of symptoms and detectability in each country shows large uncertainty. While the best estimation of the time delay in Spain is 14 days, the real value could be around 12. For Germany, for example, DD can be anywhere from just 4 days to 11 days. Using different time delays we obtain different diagnostic rates. It may be observed in Fig 5 that the percentages of diagnoses in Germany, and Switzerland are more or less constant over time. In the S1 File, we show the evolution of DD-DR for the 10 countries studied (see S1 Fig in S1 File). Assuming that this percentage remains constant to this day and that the diagnostic conditions have not changed over the last few days, we can estimate the total number of cases as of mid-

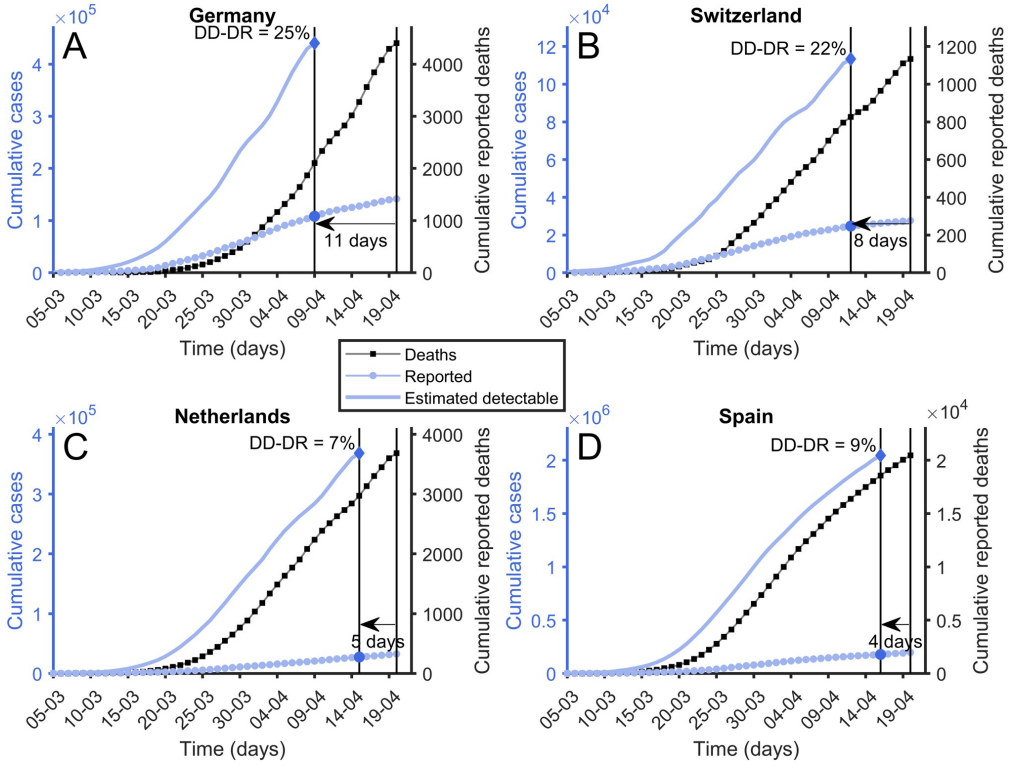

Fig 4. Delay-to-Detection Diagnostic Rate (DD-DR). Reported cumulative number of deaths (black squares), reported cumulative number of cases (blue diamonds), and estimated number of cases calculated using Eq 1 (solid blue line). To compute Delay-to-Detection Diagnostic Rate the diagnostic to death time observed in Fig 2 is used. Its value is calculated using the latest available points. (A) Germany, DtD = 10 days. (B) Switzerland, DtD = 8 days. (C) Netherlands, DtD = 5 days. (D) Spain, DtD = 4 days.

https://doi.org/10.1371/journal.pone.0243701.g004

April, 2020 to be roughly 2.3 million in Spain and close to a half million in Germany. We must notice that the ENE-COVID study [40] which surveyed the real incidence in Spain found that between April 27 and May 11, 2020 seroprevalence for the entire country was 5%, representing a total amount of people infected by COVID-19 of 2.4 million.

The table in Fig 6 shows a list of the 7D-DR as of mid-April, 2020, and the DD-DR, which seems stable, together with the associated error.

Effective Potential Growth (EPG) index for policy makers

Once the diagnostic rate is known, it is straightforward to establish a real incidence no longer affected by the presence of important differences in the time delays to diagnosis in different countries (see the table in Fig 6). The level of diagnosis and the real incidence is indeed useful for policymakers since it gives a clear general picture. However, the policy response needed to improve the diagnostic rate is conditioned, in the short-term, by the ability to increase the production of PCR kits and other diagnostic tools.

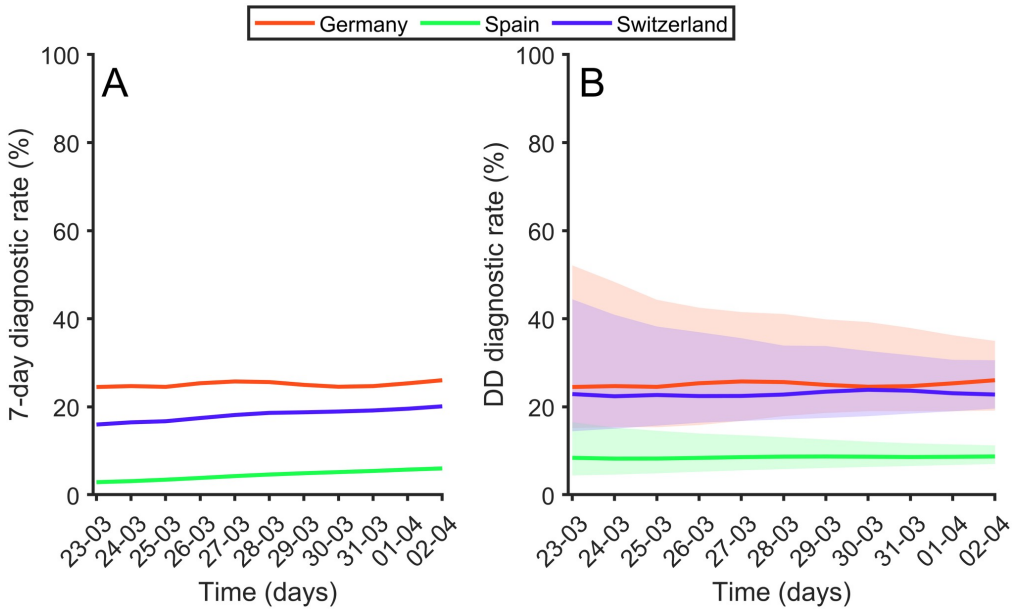

Fig 5. Diagnostic rate over time. (A) 7-Days Diagnostic rate over time for Germany (red), Spain (green), and Switzerland (blue). (B) Delay-to-Detection Diagnostic Rate over time. Thick lines are derived from the Diagnosis-to-Death time observed in Fig 2. Shaded areas represent the limits considering error bars observed in Fig 2.

https://doi.org/10.1371/journal.pone.0243701.g005

Policymakers have greater ability to immediately affect mobility patterns and social contact. In this sense, a key figure for policymakers would be to have a reliable and robust estimation of the number of infected people in each country that can propagate the disease. Providing an exact number is, right now, impossible.

We can, however, produce an index of the effective potential growth using the DD-DR and the guidelines used by the ECDC to track the epidemic. Even if the precise number of people

Country	Detection Delay			7D DR (%)	Delay to Detection Diagnostic Rate (%)			Estimated cumulative cases			Estimated attack rate /10^5 inh.		
	Value	min	max		Value	min	max	Value	min	max	Value	min	max
Belgium*	8	4	13	4%	4%	3%	6%	870000	690000	1110000	7500	6000	9600
France	10	6	14	4%	5%	4%	5%	2400000	2100000	2800000	3600	3200	4300
Germany	7	4	11	25%	25%	22%	28%	580000	510000	650000	690	610	780
Italy	14	9	14	6%	7%	6%	7%	2600000	2600000	2900000	4200	4200	4700
Netherlands	13	8	14	6%	7%	6%	8%	440000	430000	550000	2600	2500	3200
Portugal	13	9	14	19%	25%	21%	25%	82000	80000	95000	810	780	930
Spain	14	11	14	7%	9%	8%	9%	2300000	2300000	2400000	4800	4800	5100
Sweden	9	4	14	5%	6%	4%	8%	230000	180000	320000	2300	1800	3200
Switzerland	10	6	14	20%	22%	20%	23%	130000	120000	140000	1500	1400	1600
United Kingdom	14	8	14	4%	6%	4%	6%	2000000	2000000	2900000	2900	2900	4300

Fig 6. Detection Delay (DD), 7-Day Detection Rate (7D-DR), Delay-to-Detection Diagnostic Rate (DD-DR), estimated cumulative cases, and estimated attack rate. To interpret estimated cumulative cases and estimated attack rate we must take into account Detection Delay, because these are computed using the reported data. Data updated on April 20, 2020. Belgian data are biased due to reporting of unconfirmed death cases. On that date the shift was 48% [34].

https://doi.org/10.1371/journal.pone.0243701.g006

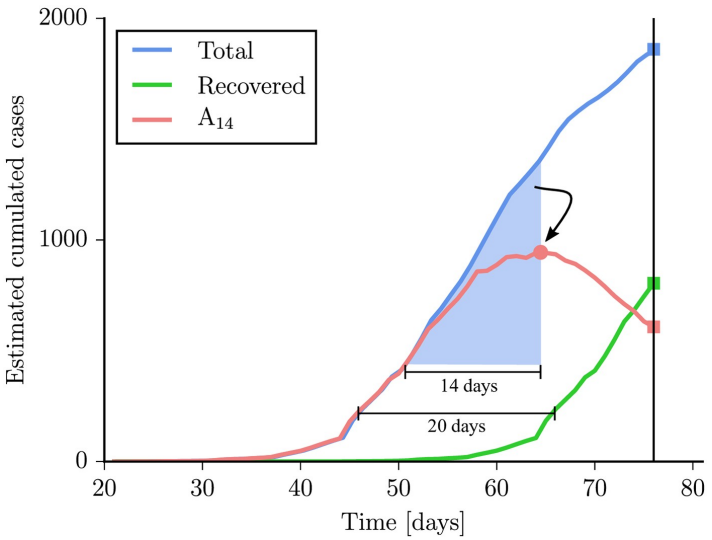

Fig 7. Schematics of the procedure to obtain incidence A_{14}, recovered and estimated cases using Germany as an example. Incidence of estimated cases (blue), contagious incidence (red), and total estimated recovered cases (green). Blue shaded area is the number of cases used to compute the estimated contagious incidence. To interpret final number of total cumulative cases, recovered cumulative cases and estimated attack rate we must take into account Detection Delay, because they are computed using the reported data. Similar figures for all countries are shown in S2 Fig in S1 File.

https://doi.org/10.1371/journal.pone.0243701.g007

with the disease were known, and the distribution of symptoms by sex and age was reported, there is no clear knowledge regarding the level of infectivity of the different types of person and symptom. For instance, it is not known for how many days an asymptomatic, pre-symptomatic or symptomatic person can transmit the disease [41–43]. Virus loads in the throat seem to be rather high across the board [44], but data on how this influences contagion is unclear. The only way to assess the situation is to use a general unbiased measure which is indicative of the potential for infection. The ECDC uses the number of newly infected people in the last 14 days [45]. We use this same criterion.

Fig 7 shows how to compute an estimation of the people that go undetected and have the potential to transmit the disease. Using the DD-DR one can compute how many undetected people were added to the infected number in the last 14 days, I_{14}. This number can only be obtained properly some days in the past, on the day that we have a typical diagnosis. After that, we would need input from new data to properly compute how many people are diagnosed. So the number I_{14} is strictly a measure of the recent past, but good enough to give a proper picture of what that the system will face in the following days.

We also consider those undetected cases which appear earlier than 14 days as recovered R_I. Note that here we use the word recovered weakly. It does not mean literally that all of them are fully recovered since most of them never fell ill to begin with, and some of them could not have neutralized tests yet, but simply that those infected and undetected more than two weeks ago do not seem to pose a serious risk.

A list of values for I_{14} and the corresponding 14-day attack rate per 10^5 inhabitants (A_{14}) is provided for each country in the table of Fig 8 with the number computed at the beginning of

Country	Estimated recovered			Estimated recovered /10^5 inh.			Active cases (last 14 days) I_{14}			Active cases (last 14 days) /10^5 inh. A_{14}			$\bar{\rho}_{(7)}$	EPG		
	Value	min	max	Value	min	max	Value	min	max	Value	min	max		Value	min	max
Belgium*	270000	210000	340000	2300	1800	2900	430000	340000	550000	3700	2900	4700	0,97	3600	2900	4500
France	960000	840000	1140000	1500	1300	1700	880000	780000	1050000	1300	1200	1600	0,72	960	850	1150
Germany	250000	220000	290000	300	260	340	190000	170000	210000	230	200	260	0,82	180	160	210
Italy	1400000	1400000	1600000	2400	2400	2700	720000	720000	810000	1200	1200	1300	0,88	1050	1050	1180
Netherlands	160000	150000	200000	910	880	1140	200000	190000	250000	1200	1100	1500	0,95	1100	1080	1400
Portugal	27000	26000	31000	260	250	300	37000	36000	42000	360	350	420	0,75	270	260	310
Spain	980000	980000	1050000	2100	2100	2200	760000	760000	810000	1600	1600	1700	0,74	940	920	1030
Sweden	64000	52000	91000	640	510	900	120000	100000	170000	1200	1000	1700	1,05	1300	1050	1800
Switzerland	69000	65000	77000	800	750	890	31000	29000	34000	360	340	400	0,70	250	240	280
United Kingdom	360000	360000	540000	530	530	800	1200000	1200000	1800000	1700	1700	2600	1,00	2000	2000	2800

Fig 8. Estimated recovered and active cases, $\bar{\rho}_{(7)}$ and EPG. $\bar{\rho}_{(7)}$ is computed using the mean value for the last three days.EPG: Effective Potential Growth described in the text. To interpret table data we must take into account Detection Delay, because they are computed using the reported data. Data updated on April 20, 2020. *Belgian data are biased due to reporting of unconfirmed death cases. On that date the shift was 48% [34].

https://doi.org/10.1371/journal.pone.0243701.g008

April, 2020. These values are currently being monitored each day for all UE countries. Having an unbiased assessment of the risk regarding the number of potential spreaders, I_{14} and A_{14}, is key to policymakers.

A_{14} alone, however, does not give a full picture of the situation. It is not the same to have 100 contagious per 10^5 inhabitants when the number of contacts is high as when the number of contacts is low. It is important to take into account the level of spreading velocity of the epidemic related to the effective reproductive number (R_t).

The effective reproductive number depends on multiple factors, from the properties of the virus itself to the number and types of contacts. Those, again, depend on different social behavior and structure such as mobility, density, and typical size of the family unit sharing a house, to name a few. The only feasible way to estimate R_t is using fits from Susceptible-Exposed-Infectious-Recovered (SEIR) models. Complex SEIR models which include spatial and contact-processes have a large number of parameters which, due to the present lack of knowledge, are unknown. This makes any estimation of R_t highly dependent on the value of other co-factors that strongly affect propagation. In essence, R_t can only be fit in very simple SEIR models where a small number of parameters are unknown and R_t can be calibrated from them.

Given the partially empirical nature of present R_t, we prefer to take a fully empirical surrogate as a quantitative evaluation of the level of infections. We define an alternative reproductive number as the number of new cases detected today divided by the number of new cases detected five days ago: N_t/N_{t-5}. However, the high fluctuations of these quantities requires that we use averaged values [9]:

$$\rho_t = \frac{\sum_{i=-n_d}^{n_d} N_{t+i}}{\sum_{i=-n_d}^{n_d} N_{t+\tau+i}}, \tag{4}$$

where N_t stands for new cases reported at day t. Specifically we use values over three days (n_d = 1) and the delay τ = 5. We take 5 days as the key delay unit since this is roughly the time at which infected people develop symptoms if they do develop them.

$$\rho_t = \frac{N_{t-1} + N_t + N_{t+1}}{N_{t-6} + N_{t-5} + N_{t-4}}. \tag{5}$$

This rate is one if the number of new cases is constant. It will be below 1 if new cases are decreasing and larger than 1 if the number of cases is increasing.

There are still clear fluctuations on a day-to-day basis of this measure ρ_t due to common delay and irregularities in reporting. In April, 2020 there was also a slight weekend effect. Due to the reduction in working hours in the information systems and biotechnological firms, the number of new cases dropped during the weekends in every single country. The backlog of cases were normally declared the following days. This fact appears in the data reported until July 2020, and it has not diminished during the epidemic. Information systems are still not required to report on a day-to-day basis. A short average of three or four days should be enough to obtain an index which is representative fast enough to the situation. However, this weekend effect requires 7-day averages to get a proper picture. We define the average of ρ_t during seven days $\bar{\rho}_{(7)}$.

$$\bar{\rho}_{(7)} = \frac{1}{7}\sum_{i=-3}^{3}\rho_{t+i}. \tag{6}$$

We propose the following day-to-day EPG index:

$$\mathrm{EPG} = \bar{\rho}_{(7)}\mathrm{A}_{14}. \tag{7}$$

EPG is just the multiplication of the growth rate of the disease $\bar{\rho}_{(7)}$ with the estimation of A_{14} both evaluated at the proper time in the recent past. The worst case scenario is one where both A_{14} and $\bar{\rho}_{(7)}$ are large. This means you had a large population with the disease and lots of spreading a few earlier. The best situation is a low value of velocity and low number of active cases. Having a large number of A_{14} with low $\bar{\rho}_{(7)}$ or a large $\bar{\rho}_{(7)}$ with low A_{14} are potentially dangerous situations. These values for the European countries we tracked on April 20, 2020, are given in the table of Fig 8. These values can be updated every day [9].

Robustness of the EPG index

In this section, we check the robustness of the EPG index against the different parameters used. In order to compute the estimated EPG, we use the lethality of the virus (1%) and the delay between the onset of symptoms and death (TtD = 18 days). Both variables are estimated based on previous works [10, 13, 23, 24, 40]. The constants related to time delay in detection and from detection to death (DD and DtD) are calculated analysing the correlations and, consequently, they cannot be tuned. Similarly, the definition of active cases as those who have been detected in the last 14 days is used according to ECDC [32]. In order to smooth the effect due to the decrease of reported data during the weekends, a seven-days average value of ρ_t is used, $\bar{\rho}_{(7)}$.

However, there are two parameters that can be modified. First, in Eq (4), we assumed $\tau = 5$ days. This is roughly the time since infected people develop symptoms if they do develop them. Given that there is a large uncertainty in this value, in the S1 File we study the sensitivity of $\bar{\rho}_{(7)}$ to changes in τ (3 and 7 days, see S1 Table in S1 File). No important differences are observed. The second parameter in Eq (4) is the number of days considered in the numerator and denominator, n_d. We take it to be $n_d = 1$ but they could certainly be more. In order to test its robustness, we compare this value with the one using $n_d = 3$ days (one week) (see S1 Table in S1 File). It is straightforward to observe that the differences are small. The same analysis is presented there for the EPG index (see S2 Table in S1 File).

We must also address how robust the EPG has been in describing the epidemic so far. This was done by considering the evolution of four key European countries: Denmark, Germany, Spain, and Sweden. Fig 9 shows that the EPG behaved exactly as expected in the description of the different situations. The evolution of the epidemic in Denmark and Germany has been

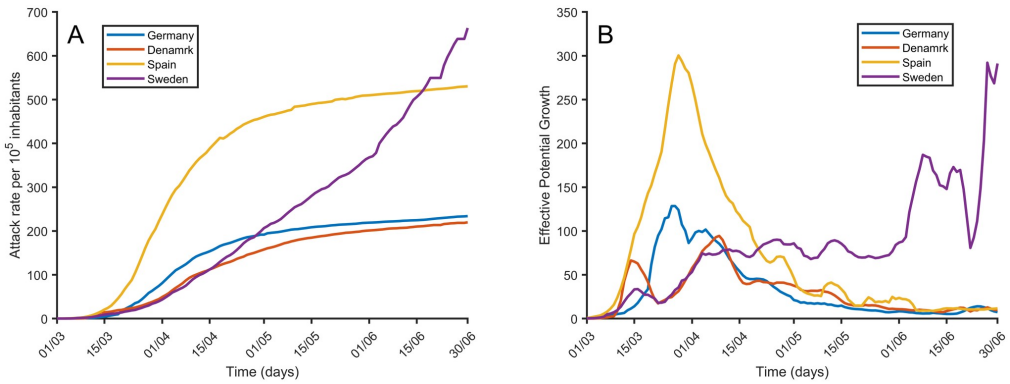

Fig 9. Attack rate and EPG for several countries. Temporal evolution of the attack rate per 10^5 inhabitants (left) and the corresponding EPG (right) for four European countries with different epidemiological dynamics.

https://doi.org/10.1371/journal.pone.0243701.g009

good and the EPG shows that this is the case. In Spain, a sharp increase in the attack rate followed by a rather fast decay is observed. The EPG (Fig 9B) increases earlier and consequently, if compared with the number of new cases, it is able to advance the troublesome situation in Spain. After the peak, it properly shows a decay showing a situation under control. On the other hand, Sweden has followed a trajectory of different stages without a sharp increase and later decay. During the later weeks depicted in Fig 9, we may observe an increase in EPG, which has been followed by an increase in cases, as expected.

We have shown that an increase of the EPG indicates correctly that the incidence during the first phase of the epidemic will increase during the following days. Additionally, it would be interesting to also analyze if the evolution of the EPG can be used as a proxy to predict secondary outbreaks once the first wave of the epidemic has abated, and what is the delay between the beginning of the outbreak and the EPG-response. European countries, as of mid-July 2020, do not present these secondary outbreaks. We need to look for a good example outside of Europe. We think the perfect example to analyze secondary outbreaks is Iran. The country has suffered a secondary and, probably, a tertiary outbreak as indicate the evolution of the number of cumulative cases shown in Fig 10A. The figure also shows that the EPG index properly predicts the new outbreaks in Iran.

However, the only way to provide quantitative values for the delay between the beginning of the outbreak and the EPG-response is to use surrogate data which properly mimics the evolution of the epidemic. In surrogate data, we can control precisely the exact moment of the initiation of the outbreak. To this end, we use Gompertz-like generated data. The evolution of the epidemic in all the countries of the world and in the different regions and states can be properly fitted by a Gompertz function [46]. Primary and secondary outbreaks can be properly fit by consecutive Gompertz functions, as shown in Fig 10B. In order to assess this delay, we simulate four different outbreaks using Gompertz surrogate structures providing different strengths for them. The expression of the cumulative cases, $C(t)$, as a function of the day t in each one of the surrogate outbreaks reads:

$$C(t) = Ke^{-\ln(\frac{K}{N_0})e^{-a(t-t_0)}}. \quad (8)$$

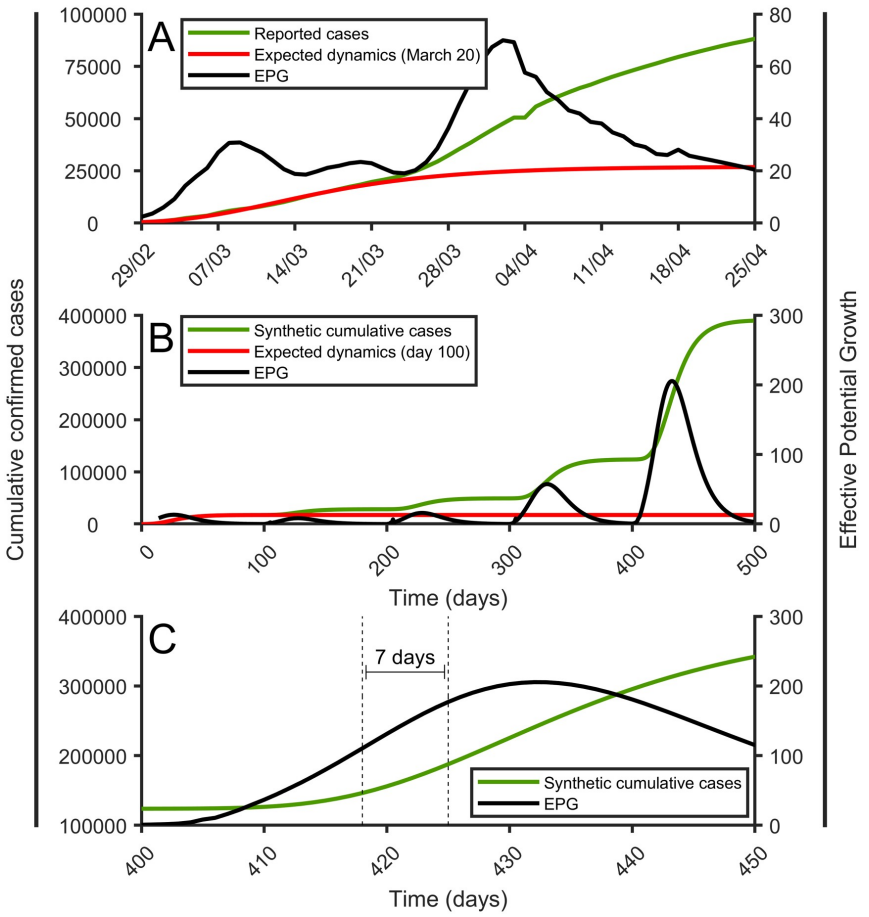

Fig 10. Evolution of the cumulative cases and the EPG. Evolution of the cumulative cases and the EPG. (A) The reported cases in The Republic of Iran (green line) is shown in comparison with the expected dynamics (red line) and with the calculated EPG (black line). (B) Same for the synthetic data. (C) Zoom of B panel between days 400 and 450.

https://doi.org/10.1371/journal.pone.0243701.g010

The different surrogate outbreaks have different parameters K and t_0, being the final number of cases of the secondary outbreak and the time where we introduce it, respectively. Fig 10B shows the four surrogate consecutive outbreaks with increasing magnitude and the associated EPG. In all of them, the EPG raises and detects the increase. If the new outbreak is small the EPG never crosses the threshold value of 100, indicating that, for the secondary outbreak, a low value of K does not produce significant increases in EPG. As long as the increase is large, EPG reaches values above 50 very rapidly. In Fig 10C we can observe the delay between the starting day of the surrogate outbreak t_0 and the peak in the evolution of the EPG. We can see how the outbreak peaks just two weeks after the outbreak. If we focus on its rapid initial increase, the outbreak can be detected clearly within 5-10 days of its onset. This is very early

considering the time needed for the epidemic to develop, with typical times scales of infection around five days or one week. This surrogate analysis shows how robust is the EPG index in detecting relevant secondary outbreaks.

Discussion

The reported number of deaths per 100,000 people is a fairly objective and relatively simple way of assessing the situation of the COVID-19 epidemic in the different countries. The complete picture must be given by a more complex analysis of other data such as the number of diagnoses per 100,000 inhabitants, distribution of these cases among regions and according to age and sex, percentages of asymptomatic and mild cases, and spreading rate of the epidemic, among others. Nevertheless, any analysis based on diagnosed cases is biased by diagnosis protocols and ratios in each country, as well as by the pool of asymptomatic cases. Moreover, any attempt to improve diagnostic percentage requires an economic, infrastructural, and logistical effort that is not always possible. In addition, this health system structure imposes a strong condition that limits the possible actions to carry out in this direction. The reported number of deaths, if uniformly and properly recorded, provides very relevant information as a first general overview. Even in countries where there is a bias in reporting death, the effort that should be made to improve these data collection is much lower than the effort needed to increase data about cases.

The assumption of a common IFR, which has been situated around 1%, allows for using the IFR as an indicator of the real incidence. Current information on IFR is still not complete, since many countries do not report distribution of death by age or sex, nor information on COVID-19 mortality outside hospitals. However, we argue that the picture that we obtain from the analysis using IFR is closer to reality than the one provided by the pure analysis of reported cases. In particular, this analysis allows for: (1) establishing an order of magnitude of real cases and diagnostic percentage, (2) assessing an effective potential growth index to evaluate the risk, and (3) obtaining an order of magnitude of recovered people that could be potentially immunized in the short-term.

In Europe, absolute case ranking has been lead by Italy (until April 4, 2020) and Spain (since then). On April 20, 2020, Spain was at the level of 196,000 reported cases while Italy was reporting 179,000. They were followed by Germany (142,000), the United Kingdom (120,000), France (113,000), and Belgium (38,000). If we estimate the cases that should have been diagnosed by that time, the ranking is led by Italy (2,600,000) and followed by France (2,400,000), Spain (2,300,000), the United Kingdom (2,000,000), Belgium* [33] (870,000), and Germany (580,000). Thus, differences in diagnostic rate are absolutely significant when analyzing the global situation in Europe. In April 2020, countries like Germany, Portugal, and Switzerland would be diagnosing around 25% of cases, while Belgium, France, Sweden, and the United Kingdom would be at the level of 5%.

Assessing the risk of countries to enter or remain in the epidemic growth phase is essential. In this sense, the EPG index is a valuable tool for policy makers. A high EPG in the situation where there is a high growth rate of the epidemic and a large number of active cases is a clear situation of danger, while a very low EPG because both the number of recently infected and the spread velocity are low is clearly a controlled situation. In intermediate situations, EPG informs whether the growth rate is too high for the number of infected at hand. Even if $\bar{\rho}_{(7)} < 1$ and the epidemic seems to be under control because new cases are decreasing, intermediate EPG values informs the policy maker that reopening can have a very important cost in the form of secondary outbreaks and waves of infection. A rather large EPG with low $\bar{\rho}_{(7)}$ is a

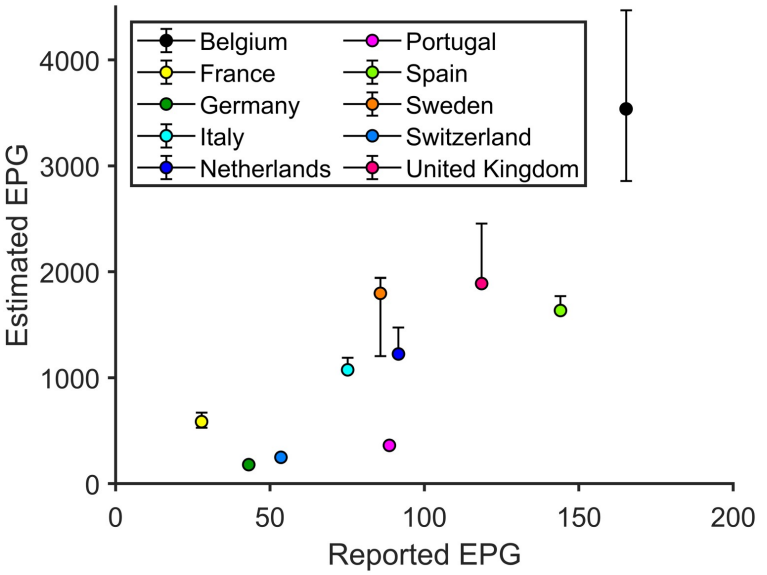

Fig 11. Reported EPG vs estimated real EPG. Several European countries in terms of the EPG computed using the reported data on the attack rate vs the EPG using our estimation of the real attack rate. The order of the different countries should be done from right to left (for the reported state of the index) and from top to bottom (for the estimated value of the index). We observe how the comparative situation of the different countries changes as of April 20, 2020. * Belgian data are biased due to reporting of unconfirmed death cases. On that date the shift was 48% [34].

https://doi.org/10.1371/journal.pone.0243701.g011

situation in which the number of spreaders is potentially very high and increasing the number of contacts carries a large risk. Therefore, EPG is a very informative index that is also very robust.

Although $\bar{\rho}_{(7)}$ is independent of the diagnostic rate, reported I_{14} directly depends on the level of diagnosis. Thus, if EPG is evaluated by using the reported data (reported EPG), it can provide an erroneous picture of the situation. Based on reported EPG, the worst situation in Europe on April 20, 2020 would be that of Belgium, followed by Spain, the United Kingdom, the Netherlands and Portugal. If risk is evaluated with the EPG calculated using the estimated cases (estimated EPG), the highest value would still correspond to Belgium, but followed by Sweden, the United Kingdom, Spain, the Netherlands, and Italy. Portugal is in a much better position than its reported data suggest. Actually, countries with similar reported EPG like that of Portugal, and Netherlands have, in fact, a totally different estimated EPG, with the last country at significantly higher risk than the other nine. Fig 11 shows visually how reported and estimated EPG compare.

We have shown in the Methods section that the basis for obtaining estimated I_{14} and A_{14} is not biased due to demographic differences and, at present, there is no indication that it is biased due to a different way of accounting for the cumulative death toll of the epidemic. There is also no indication that comorbidity factors are significantly different in different countries or that IFR is higher in some countries given that ICU units and hospitals are not available for people that might need them, at least to date. If this were the case, under any scenario where the situation occurs, the epidemic in that country would have a number of cases,

10 Empirical model for short-time prediction of COVID-19 spreading

10.1 Summary

When the COVID-19 epidemic arrived in Europe, lockdowns were the only solution that governments found to stop the growth of COVID-19 cases. This occurred because they were not prepared and because of the unknowns that we had about this disease. The main goal of this paper is to find a model to predict short-term evolution pandemic.

Initially, we had no clue of the main parameters that marked epidemic evolution. We did not know what the most efficient measures were or the relation between each measure and its epidemiological effect. In this article we build an empirical model that allows us to make predictions without having all this knowledge. The drawback is that this model can perform reliable predictions when conditions are maintained, but it can not anticipate significant changes in the dynamics. When there is a sudden change, the model requires some days to capture the new trend.

The Gompertz model was originally used to describe human mortality curves [72], but it has been used in other areas like the growth of cultures [86], and to forecast an epidemics [24]. It is a two parameter model that is very similar to the logistic one but without its symmetric behaviour, i.e., the decrease in new cases is slower than the initial growth.

Firstly, we demonstrate that the Gompertz curve fits the Chinese territories where the epidemic's initial wave had nearly finished very well. The Gompertz curve was adjusted to the cumulative counts of cases. With consecutive fittings, we could see how the fitting adjusted to the available data each day. This allows us to see how parameters values converge on their final values. Extending the Gompertz model to the future, we could make predictions and we evaluated their performance and where errors lay.

This consecutive prediction allowed us to see that, at the initial stage of the epidemic, we could make good short-term predictions but we could not estimate the final number of cases, for example. In fact, another parameter that can not be anticipated is the epidemic peak, but we can affirm the epidemic peak has been passed some days afterwards. However, the day that 90% of the final cases have been reached can be

predicted several days beforehand. The final number of cases can also be predicted after reaching the epidemic peak with a good interval error.

Short-term prediction results are fine and the number of cases are normally found within the confidence interval. In fact, a mean error for a 5-day predictions is 8% and in 70% of cases its value is found between the confidence intervals. These predictions are refined by adding an extra weight at the last three available data points to better detect sudden changes.

10.2 Main contributions

The main contributions of this article are:

- An empirical computational model that describes the evolution of the COVID-19 epidemic is built. The Gompertz model is fitted to the cumulative and new COVID-19 cases in different Chinese territories (where the epidemic was in a mitigation stage) and to the new growing epidemic in European countries.
- Best parameters set (it is a two-parameter model: a and K) is determined and adjusted for each of the studied countries. Extending the fitting to the future allows us to perform short term predictions for each of the desired countries and territories.
- Short-term predictions errors are determined for the following 5 days. Prediction interval success probability is also assessed. By observing the temporal evolution of parameter fit values and their intervals, we can determine how many days in advance the different model parameters could be predicted (K, t_p, t_90 ...).
- Cumulative COVID-19 case forecast in European countries for the coming days are then made.

10.3 Application and further work

This method has been used in BIOCOM-SC COVID-19 reports for European Commission DG-CONNECT [38]. It has also been used in daily reports deliverd to Catalan health authorities: Agència de Qualitat i Avaluació Sanitàries de Catalunya (AQuAS) and Instut Català de la Salut (ICS). These predictions were used to compare and refine a neural network predictor develped by Facebook Artificial Intelligence [98, 104]. We sent scripts to use these predictions to several research groups in

different countries: Mexico, Peru, Brazil, Argentina, United States of America, Zimbabwe, Senegal, Nigeria, Sudan, and Ecuador. In fact, the Brazilian group translated our codes to python and R and published them in a public GitHub repository [6]. They have also been used to submit periodic predictions to the European Covid-19 Forecast Hub [63].

This method is in constant revision to adapt to the new epidemiological situations. First we added a constant, non-zero, starting point to make predictions during the second, third and consecutive waves. One problem that scientists face when forecasting COVID-19 cases is that there is not a homogeneous reporting pattern. In fact, there is a well-described weekend effect [44, 39, 40] and other effects due to holidays that impede performing predictions. These problems were solved by adapting the predictions to the reporting pattern of each country [41]. A manuscript to explain this in detail in in preparation.

10.4 Article

The full article can be read online at:
Català M, Alonso S, Alvarez-Lacalle E, López D, Cardona PJ, et al. (2020) Empirical model for short-time prediction of COVID-19 spreading. *PLOS Computational Biology.* 16(12): e1008431. `https://doi.org/10.1371/journal.pcbi.1008431`

Supplementary figure of evolution and prediction of COVID.19 cases on 3 May 2020 for European countries can be seen at: `https://doi.org/10.1371/journal.pcbi.1008431.s001`

Supplementary figure of estimation evolution of a parameter for European countries up to 3 May 2020 can be seen at: `https://doi.org/10.1371/journal.pcbi.1008431.s002`

Supplementary figure of estimation evolution of K parameter for European countries up to 3 May 2020 can be seen at: `https://doi.org/10.1371/journal.pcbi.1008431.s003`

Supplementary figure of estimation evolution of t_p parameter for European countries up to 3 May 2020 can be seen at: `https://doi.org/10.1371/journal.pcbi.1008431.s004`

Supplementary figure of estimation evolution of t_90 parameter for European countries up to 3 May 2020 can be seen at: `https://doi.`

org/10.1371/journal.pcbi.1008431.s005

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Empirical model for short-time prediction of COVID-19 spreading

Martí Català[1,2], Sergio Alonso[2]*, Enrique Alvarez-Lacalle[2], Daniel López[2], Pere-Joan Cardona[1,3,4], Clara Prats[1,2]

1 Comparative Medicine and Bioimage Centre of Catalonia (CMCiB), Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Catalonia, Spain, **2** Department of Physics, Universitat Politècnica de Catalunya (UPC-BarcelonaTech), Barcelona, Catalonia, Spain, **3** Experimental Tuberculosis Unit, Fundació Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Catalonia, Spain, **4** Centro de Investigación Biomédica en Red de Enfermedades Respiratorias, Madrid, Spain

* s.alonso@upc.edu

OPEN ACCESS

Citation: Català M, Alonso S, Alvarez-Lacalle E, López D, Cardona P-J, Prats C (2020) Empirical model for short-time prediction of COVID-19 spreading. PLoS Comput Biol 16(12): e1008431. https://doi.org/10.1371/journal.pcbi.1008431

Editor: Benjamin Muir Althouse, Institute for Disease Modeling, UNITED STATES

Received: May 8, 2020

Accepted: October 9, 2020

Published: December 9, 2020

Data Availability Statement: The data employed had been published daily by WHO and European Centre for Disease Prevention and Control (ECDC) and is available at https://github.com/catalamarti/Gompertz_Catala2020/blob/main/Data_ECDC.xlsx.

Funding: CP, PJC and MC received funding from La Caixa Foundation (ID 100010434), under agreement LCF/PR/GN17/50300003; PJC received funding from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Grup Unitat de Tuberculosi Experimental, 2017-SGR-500; CP, DL, SA, MC received funding from Ministerio de

Abstract

The appearance and fast spreading of Covid-19 took the international community by surprise. Collaboration between researchers, public health workers, and politicians has been established to deal with the epidemic. One important contribution from researchers in epidemiology is the analysis of trends so that both the current state and short-term future trends can be carefully evaluated. Gompertz model has been shown to correctly describe the dynamics of cumulative confirmed cases, since it is characterized by a decrease in growth rate showing the effect of control measures. Thus, it provides a way to systematically quantify the Covid-19 spreading velocity and it allows short-term predictions and longer-term estimations. This model has been employed to fit the cumulative cases of Covid-19 from several European countries. Results show that there are systematic differences in spreading velocity among countries. The model predictions provide a reliable picture of the short-term evolution in countries that are in the initial stages of the Covid-19 outbreak, and may permit researchers to uncover some characteristics of the long-term evolution. These predictions can also be generalized to calculate short-term hospital and intensive care units (ICU) requirements.

Author summary

Covid-19 has brought the international scientific community into the eye of a storm. Collaboration between researchers, public health workers, and politicians is essential to deal with this challenge. One of the pieces of the puzzle is analysis of epidemiological trends so that both the current and immediate future situation can be carefully evaluated. For this reason we have employed a daily generic growing function to describe the cumulative cases of Covid-19 in several countries and regions around the world, and particularly the European countries during the Covid-19 outbreak. Our model is completely empirical, meaning it relies solely on the daily data update of new cases and does not require assumptions to make predictions. In this manuscript, we detail the methods employed

Ciencia, Innovación y Universidades and FEDER, with the project PGC2018-095456-B-I00. This work has been also partially funded by the European Commission - Directorate-General for Communications Networks, Content and Technology through the contract LC-01485746 to CP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and the degree of confidence we have obtained during this process. We obtain predictions with a success greater than 90%, which means that around 90% of the value of the reported cases are inside the prediction intervals. This can be used for other researchers collaborating with and advising health institutions around the world during the Covid-19 outbreak or any other epidemic that follows the same pattern. We hope it may help facilitate policy decisions, the review of in-place confinement measures, and the development of new protocols.

Introduction

A disease outbreak is always a challenge for public health control systems. When the outbreak is caused by a new agent able to cause a pandemic, the challenge is even greater and should involve the whole research community as well. Globalization plays a double role in this context; on the one hand, it increases the risk of the outbreak evolving towards a pandemic, while on the other, the sharing of data and strategies increases the likelihood of controling it. The new SARS-CoV-2 virus (severe acute respiratory syndrome coronavirus 2) has put the international community at the brink of a global disaster. National and local governments are working with public health agencies hand in hand to slow down, and eventually control, the spread of Covid-19 [1].

Daily availability of data about confirmed cases of Covid-19 in different regions is a unique opportunity for basic scientists to contribute to its control by carefully analyzing trends. In particular, mathematical models are widespread, as are consolidated tools to extract valuable information from the reported data on Covid-19 and help making predictions [2]. Classic SIR and SEIR models (i.e., compartment models that divide a population into Susceptible, Exposed, Infectious and Recovered) are being currently employed to evaluate and predict the spreading of the epidemic episodes [3]. They were employed in the description of the Ebola epidemic in 1995 [4] and 2014 [5] and in the more recent SARS epidemic of 2003 [6], among others. After the SARS epidemic in 2003, in order to account for the control efforts of governments, some modifications were introduced into the SEIR model to evaluate control measures [7, 8]. Furthermore, the analysis of SEIR models has been used for the modeling of Covid-19's spread in China in a effort to fit the characteristic values [9, 10]. However, during the development of the epidemic, government measures are the key drivers of the epidemic. The evolution of the disease is completely different depending on the strength and type of restrictions on mobility and social life that governments implement. The evolution of the disease in a situation where there is a total lockdown is very different from a situation where only specific restrictions to mobility apply, such as forbidding large gatherings. Similarly, the evolution is different depending on the nature of the policy initiatives. Closure of schools or bars affects the evolution differently than closure of nightlife venues. Simple SIR and SEIR models are not designed to deal with this type of situation where the network of contacts and its changes due to policy are key. SIR and SEIR models deal properly with epidemics where the key element of the evolution is the the total number of susceptible population. Its reduction, as the epidemic advances, gives the characteristic peak-like evolution. In the case of Covid-19, the total number of susceptible is not important because cumulative cases in the countries are far from achieving herd immunity [11].

There is, however, another approach based on the phenomenological comparison of the curve of cumulative cases with a typical function for growing processes. Evaluating the curve during a window of days before a particular day t allows prediction of the future short-time

behavior tendency at time $t + \Delta t$ [12]. In fact, the use of a growing function has some important advantages. Typically, the first growing function chosen is the Verhulst equation [13] which is the solution of the logistic population model and its generalization [14, 15], or the Richards model [16], which has been employed in several epidemics of smallpox, influenza, and Ebola, among others [14]. Some of these dynamic phenomenological growth models to study epidemic outbreaks have been compared in the initial phases of the Covid-19 epidemic for short-term forecasting [17].

A similar growth model is the Gompertz function [18] where the main difference is the replacement of the saturation of the growing factor, linear for the Verhulst equation and nonlinear for the Richards, and generalized Verhulst model, by an exponential decrease. These functions are similar and they have been used in the description of epidemics and in particular for studying different epidemic episodes [19, 20]. While the logistic equation produces a symmetric bell-shaped function for new cases, the Gompertz model gives rise to an asymmetric function with fast growth of new cases combined with a slow decrease, which is closer to the distribution of new cases observed in different countries during some epidemics. In this manuscript we demonstrate that the asymmetric nature of the Gompertz model is the proper framework to study epidemics in which control measures are at the heart of the evolution, since it captures the dynamic nature of the variation due to social distance measures.

Here, we employ the Gompertz growing function to analyze the dynamics of the spreading of Covid-19 in 28 European countries to make short-time predictions of the new cases for successive days. We forecast the dynamics of the pandemic in a similar fashion to the forecasting done previously with the Verhulst equation and the Richards model for Ebola epidemics [21]. The methodology and the results discussed here were employed for the writing of daily reports [22] at the very beginning of the epidemics. Later on, similar methodologies were employed to fit worldwide data [23, 24], and the data in particular countries like Mexico [25] and Brazil [26], among others. We have also applied similar methodology for the prediction of cases for hospitals and intensive care units (ICUs).

It is important to note that we forecast the dynamics of the pandemic using a phenomenological model, obtaining short-term predictions for daily new cases with over 90 percent success (see below). These data may be useful for public health policy makers and they are easily reproducible by scientists all over the world.

Materials and methods

After a short note about data acquisition, we describe the function employed for the fitting of the data and then describe the evaluation of the errors associated with these calculations.

Data acquisition

All the data employed in this manuscript have been downloaded from public repositories of the European Centre for Disease Prevention and Control (ECDC). The data contain the daily list of cumulative cases for all the countries of the world reporting the data and it is a fully open source [27], originally from [28]. Similar data are supplied by the World Health Organization (WHO) [29].

Short review of Gompertz equation

We employ the Gompertz model for growing processes to model the cumulative cases of Covid-19. The equation was originally proposed as a means to explain human mortality curves [18], and it has been further employed in the description of growth processes, for example,

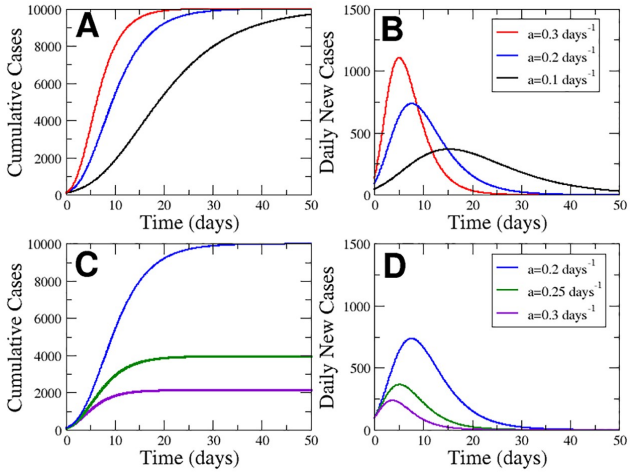

Fig 1. Properties of Gompertz function. Evolution of the cumulative cases (A) and new cases (B) keeping $K = 10^4$ for three different values of a. Evolution of cumulative cases (C) and new cases (D) keeping $\mu_0 = 0.92$ for three different values of a.

https://doi.org/10.1371/journal.pcbi.1008431.g001

bacterial colonies [30] and tumors [31]. The Gompertz equation reads:

$$N(t) = Ke^{-\ln(\frac{K}{N_0})e^{-at}} \tag{1}$$

where the parameter K corresponds to the final number of cases, N_0 is the initial number of cases for the definition of the origin of time, and parameter a is the rate of decrease in the initially exponential growth; see curves in Fig 1A for different values of a. For the beginning of the epidemic, corresponding to $t \to 0$, the Eq (1) reduces to an initial exponential growth $N = N_0 e^{\mu_0 t}$ with rate $\mu_0 = a \ln (K/N_0)$. After time t_p the growth flattens asymptotically to the final value given by the saturation parameter K. To compare with the cumulative cases of Covid-19 we begin to measure above 100 cases ($N_0 = 100$). The exponential rate μ_0 provides us with the relation between the parameters K and a.

In addition, the Gompetz function can be interpreted as the solution for the next couple of ordinary differential equations:

$$\frac{dN}{dt} = \mu N, \qquad \frac{d\mu}{dt} = -a\mu; \tag{2}$$

which corresponds, respectively, to an exponential growth with a growing rate μ which exponentially decreases with rate a.

The Gompertz function shows the cumulative cases. Therefore the temporal derivative of the cumulative cases is basically the new cases. Performing the temporal derivative we obtain:

$$N_n = \frac{dN}{dt} = aKe^{-\ln(\frac{K}{N_0})e^{-at}}\left(\ln\left(\frac{K}{N_0}\right)e^{-at}\right); \tag{3}$$

the dynamics of which as as function of time are plotted in Fig 1B.

Fixing the total values of cases ($K = 10^4$) we can study the effect of a rapid decay of the growing rate, related to a large value for a with a slower decrease, determined by a low value for a.

See Fig 1A for a visual inspection of the effect of this parameter a. The increase in the parameter a produces a delay in the growing process and delay of the peak, see Fig 1B, where the area of the curve is constant because of the conservation of the final value K. However, in Fig 1C we fix the initial exponential growth determined by μ_0 and increase the parameter a, which decreases the final value of total cases. The amplitude of the peak is decreased by the increase in the rate a when the initial growth is fixed, see Fig 1D.

We see a maximum of new cases in Fig 1, for which the inflection point (t_p) can be calculated:

$$t_p = \frac{1}{a}\ln\left(\ln\left(\frac{K}{N_0}\right)\right) = \frac{1}{a}\ln\left(\frac{\mu_0}{a}\right); \tag{4}$$

and we can also estimate the time necessary to arrive at 90% of the total value of cases K:

$$t_{90} = \frac{-1}{a}\ln\left(\frac{-\ln 0.9}{\ln(K/N_0)}\right). \tag{5}$$

The last two expresions clearly mark the effect of the parameter a. The larger the value of a, the faster the appearance of the peak and the arrival at 90% of cases, see Fig 1B and 1D.

Evaluation and propagation of errors

The fitting of the Gompertz function to the data is done with a matlab routine using the minimum least squares method [32]. This method allows for evaluation of the set of model parameters that provide the best fit for the Gompertz model to the data. Furthermore, the method also provides the error associated with the values of the fitting constants. The performance of the fitting can be evaluated with the statistical parameter R^2, available from the procedure of the calculation of the fitting.

We employ the explicit values of the fitted parameters to make our predictions. The propagation of the uncertainty or error in the calculation of the predictions can be done using the classical methods of propagation of errors [32]. In short, if we have a quantity U which depends on two magnitudes $U = U(a, b)$ and these magnitudes have their uncertainties $a \pm \delta a$ and $b \pm \delta b$, if we assume the quantities are uncorrelated we can calculate the uncertainty of the new quantity as:

$$\delta U = \sqrt{\left(\frac{\partial U}{\partial a}\right)^2 \delta a^2 + \left(\frac{\partial U}{\partial b}\right)^2 \delta b^2}, \tag{6}$$

expression which is employed for example for the calculation of the time to peak; see Eq (4) and for the calculation of the time to reach the 90% of the expected value of K. For example, we calculate the dependence of the error in t_p on the parameters a and K, see Eq (4):

$$\delta t_p = \sqrt{\left(\frac{ln\left(\ln\left(\frac{K}{N_0}\right)\right)}{a^2}\right)^2 \delta a^2 + \left(\frac{K}{aln\left(\frac{K}{N_0}\right)}\right)^2 \delta K^2}; \tag{7}$$

a similar calculation can be made for the error of t_{90}, see Eq (5).

Results

We make some predictions using the Gompertz function to fit the cumulative cases of Covid-19 in different countries where the epidemic was developed enough in April, 2020. Next, we show such predictions and the main applications of the Gompertz model for the characterization of the epidemic.

Gompertz model fits the number of cases for recovered regions

Gompertz model [33] correctly describes the trend of the cumulative confirmed cases as seen in Fig 2 where the values of the statistical measure R^2 are close to 1. We perform a systematic analysis of the dynamics of the cumulative cases of Covid-19 in different regions in China where the spreading of the epidemic finished; see for example the three regions shown in Fig 2 where the Gompertz function has been fitted. Note, however, that the fit in Hubei is divided into two regions because of a change in the protocol for reporting cases. The new cases are also fitted with relatively large R^2 values; see three panels below in Fig 2, with the function derived from the Gompertz model; see Eq (3). To fit the Gompertz function to the data we obtain the values of the fitting parameters a and K, which accompany the corresponding panels in Fig 2.

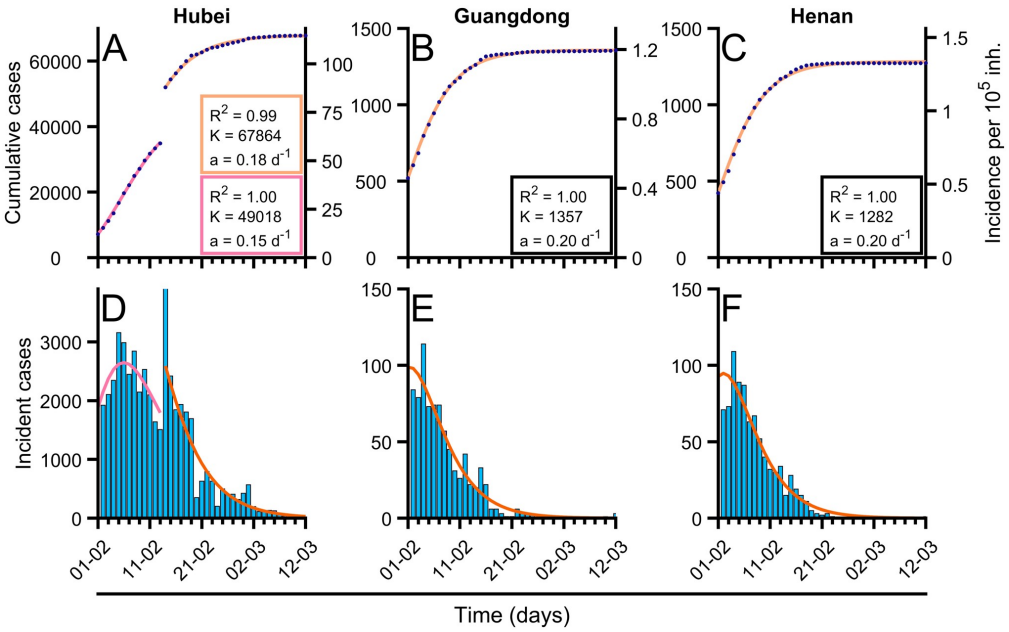

Fig 2. Fitting of Gompertz function to the cumulative confirmed cases of Covid-19 in different countries. Fitting of Gompertz function to the cumulative confirmed cases of Covid-19 in different countries. (A-C) Evolution of total confirmed cases in different regions of China (blue dots) and fitted Gompertz function in each region (orange solid line). (D-F) Evolution of new cases in different regions of China (blue bars) and fitted Gompertz function in each region (orange solid line), with $R^2 = 0.65$, $R^2 = 0.72$ (Hubei), $R^2 = 0.94$ (Guangdong), $R^2 = 0.94$ (Henan). In the case of Hubei (A, D), as there was a sudden change in reporting criterion there were two fitted Gompertz adjustments: pre-change (pink solid line) and post-change (solid orange line). The obtained values of parameter a (related with growth rate), K (final number of cases), and mean-squared error (R^2) are shown for each of the fittings. Data were updated on March 5, 2020 from [29].

https://doi.org/10.1371/journal.pcbi.1008431.g002

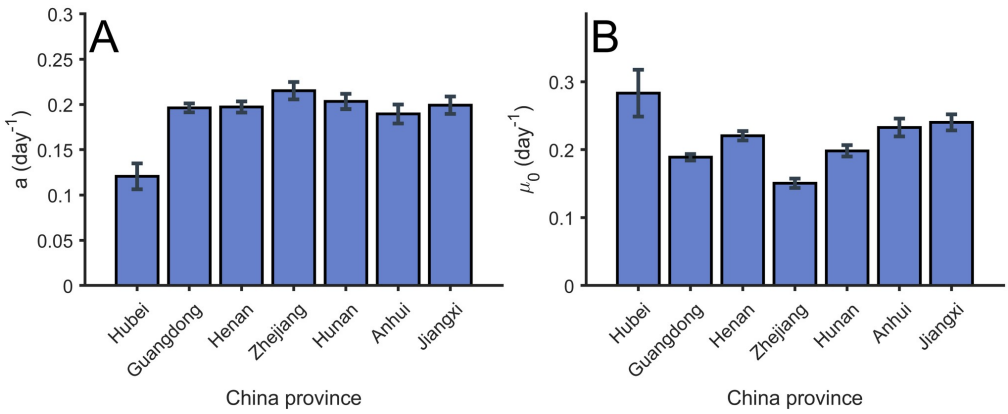

Fig 3. Values of parameter a and μ_0 for the Gompertz function in different regions in China. (A) Value of parameter a obtained from the fitting of the total confirmed cases. (B) Value of parameter μ_0 obtained from the fitting of the total confirmed cases. Error bars parameters with confidence intervals of level $\alpha = 0.01$.

https://doi.org/10.1371/journal.pcbi.1008431.g003

Let us focus now on this classification according to control measures. We show in Fig 3 the values resulting for the fitting of the Gompertz function to the data from several regions in China. Assuming that the measures of control taken in China were considered very restrictive, we can assume that the values obtained in these regions, and shown in Fig 3A, are the upper limit of the parameter for other countries. The actual value obtained is around a = 0.2 days^{-1}.

Furthermore, we can evaluate the value of parameter μ_0 for the initial exponential growth of the different regions; see details in Fig 3B. We obtain similar quantities in all the regions in China and it provides information about the growing rate of the epidemic in China, the value of which is, in these cases, similar to the decreasing rate a calculated above.

Short-term predictions obtained from Gompertz model

Although understanding of the epidemic from the final picture of the dynamics is a valuable result for the treatment of future epidemics, the main goal of the modeling of epidemics is the actual possibility of prediction of the behavior during the incidence of the epidemic. We have used the Gompertz model during the epidemic episode of Covid-19 in several countries in Europe.

First, we evaluated the predictions with the data obtained in the different regions in China to estimate the error of the fitting procedure of the Gompertz function before saturation of the number of cases. We began with the first day after 100 cumulative cases of Covid-19 and we successively fit a Gompertz function to the previous values of cumulative cases to estimate the values of parameters a and μ which permits estimation of the values for the cases for the next days. In Fig 4A, 4B and 4C), we show the fitting of the Gompertz function to the values of cumulative cases at three different times. The fittings of the function at different times differ with the final values of the total function shown in Fig 2D and 2E and therefore the values of the three fittings produce different values of parameters a and μ_0. However, the evolution of the values converges to the global fitting of the function to the whole set of data, see Fig 3D, 3E and 3F.

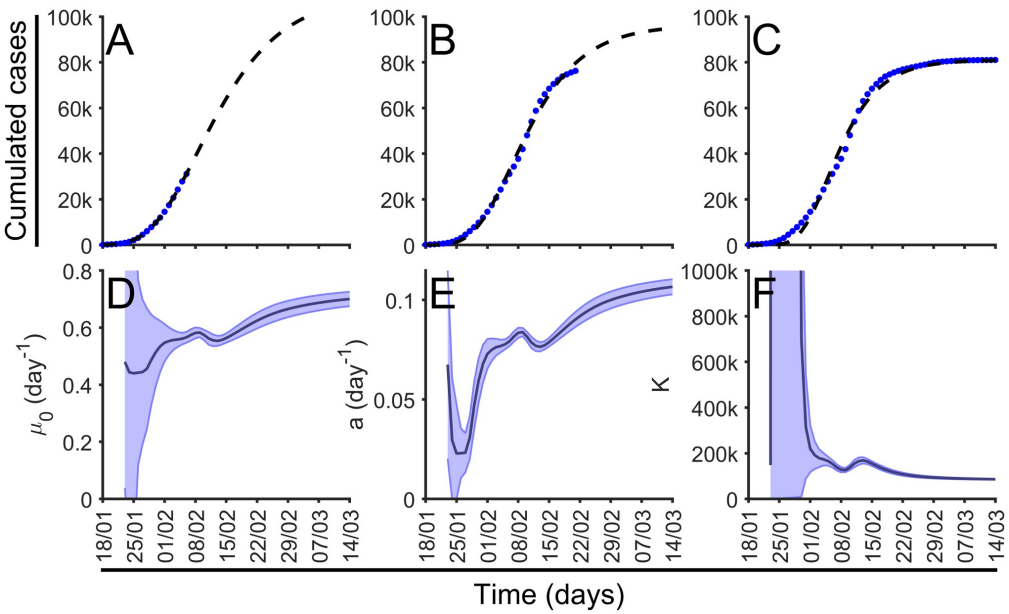

Fig 4. Dynamical fitting of Gompertz function and parameters evolution. (A-C) Gompertz fitting for China at three different time points, 7 February, 20 February and 14 March. Number of cumulative cases (blue dots) shown together with the function fitted (black dash line). (D-F) Dynamic calculation of parameters μ_0, a, and K in dark blue; light blue mark error bars parameters with confidence intervals of level $\alpha = 0.01$.

https://doi.org/10.1371/journal.pcbi.1008431.g004

Such large variations on the parameter fittings show clearly that long-time predictions are complicated. However, we can perform short-time predictions for the number of new cases if we extrapolate the Gompertz function to the near future with the updated values of a and μ_0 for the cumulative cases. We systematically extrapolated the new cases for each temporal data of the series of cumulative cases of Covid-19 in the different regions in China and obtained a successful agreement of the predictions with the actual data for the whole series; see below for more extensive results taking into account a larger number of countries.

Short-term predictions can be applied to ongoing epidemics

The epidemic is still spreading throughout Europe and we have been fitting the Gompertz function to the total cumulative cases for two months (March and April 2020). Most of the countries had already arrived at the saturation stage and the fitting of the function allows evaluation of the control measures. See the examples in Fig 5, where a Gompertz function satisfactorily fits the existing data. Note that Gompertz function is able to fit countries at different epidemiological phases. We systematically assessed short-time predictions for all European countries, the United Kingdom, Norway, and Switzerland every day from March 17th [34], as well as for Spanish and Italian regions [35].

Typically, the evolution of confirmed cases shows a biphasic behavior: an initial lag phase where no significant increase in the incidence is observed, which would correspond to the period where most of the cases are imported, followed by a subsequent phase where growth is evident, which would be a reflection of triggering local transmission. Gompertz model is fitted

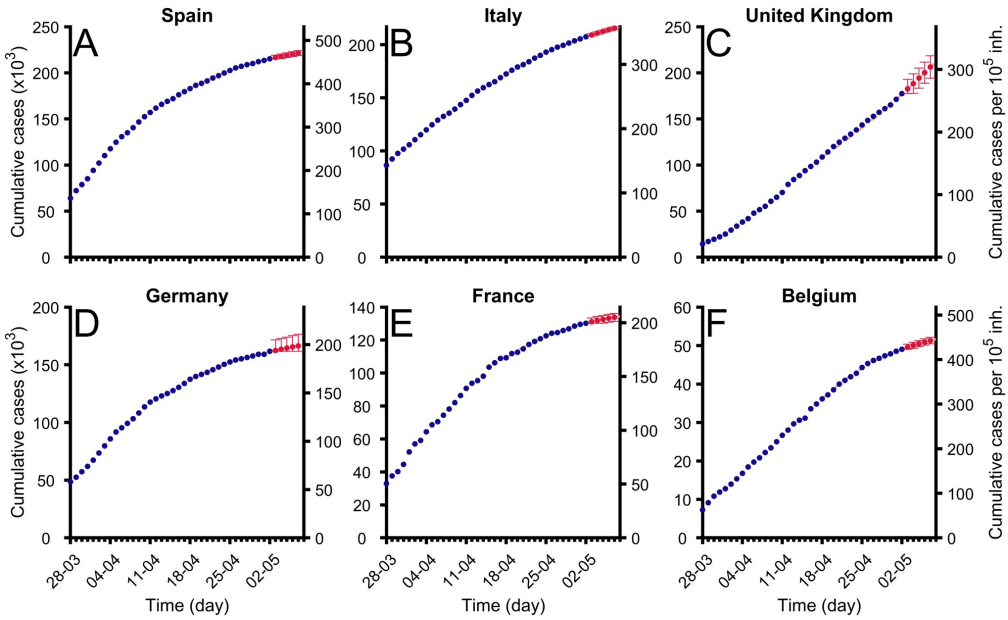

Fig 5. Fitting of Gompertz function to cumulative cases in some countries in Europe. Evolution of total confirmed cases in different regions (blue dots) and fitted Gompertz function in each region (orange dashed line). Red points show predictions for next 5 days and error bars marks their confidence interval levels $\alpha = 0.01$. Data were updated on April 9, 2020 from [29]. (A) Spain (B) Italy (C) Germany (D) France (E) United Kingdom, and (F) Belgium.

https://doi.org/10.1371/journal.pcbi.1008431.g005

to the later phase, i.e., it is applied from the moment when a clear increase in confirmed cases is observed, typically above 100 cases to avoid the evaluation of the beginning of the epidemic dominated by the importing of cases from other zones.

As an example of these predictions, we refer the reader to Table 1. The cases correspond to the evolution of the values of the cumulative cases up to April 29, 2020. We show the predictions for some countries in Europe of the algorithm based on the Gompertz function, for the next 1, 3, and 5 days. The rate of success in this example is representative for the algorithm; see following section.

We fit the function over time to be able to predict the evolution of the cumulative cases to generate some useful information which may help political institutions to adopt appropriate control measures; see supplementary S1 Fig for approximations of a selection of countries in Europe. Such curves are based on the calculation of the values of a; see supplementary S2 Fig, and K, see supplementary S3 Fig, in the selection of countries.

Evaluation of the errors in the short-term predictions obtained with the Gompertz model

To evaluate the quality of the predictions we systematically ran the prediction routines along the past, for all the days of the spreading of Covid-19 in all countries with more than 1000 cases as of April 11, 2020. We compared the prediction with the actual number of cases to give rise to two different indexes: first, the average relative error of the prediction with the real

Table 1. Short-term predictions on April 29, 2020 with Gompertz model. Countries were sorted by number of reported cases. The top 10 countries in terms of cases cases were chosen from among the UE+EFTA+UK. Predictions are the number of cases at April 30, May 2, and May 4, respectively; lower and upper bounds can be seen inside brackets. In bold reported cases that were inside prediction intervals. K is the predicted final number of cases.

Countries	Cases	April 30		May 2	
		Prediction	Reported	Prediction	Reported
Spain	213942	215365 [213942-220067]	**215183**	217465 [213942-222263]	**217804**
Italy	201505	203403 [201505-206418]	**203591**	206856 [203761-209951]	**207428**
United Kingdom	161145	165268 [162977-167560]	**165221**	172953 [170582 -175324]	177454
Germany	157641	158753 [157641-160596]	**159119**	160718 [158837-162598]	**161703**
France	126835	127941 [126835-130010]	**128442**	129882 [127761-132004]	**130185**
Belgium	49227	49698 [49227-51501]	**49741**	50616 [49227-52459]	**50565**
Netherlands	38416	38889 [38416-41086]	**38802**	39462 [38416-41696]	**39791**
Switzerland	29181	29279 [29181-29545]	**29324**	29433 [29181-29703]	**29622**
Portugal	24324	24654 [24324-25941]	**24692**	25113 [24324-26427]	**25351**
Ireland	19877	20253 [19877-21523]	**20253**	20791 [19877-22087]	**20833**

Countries	Cases	May 4		Parameter K	
		Prediction	Reported		
Spain	213942	219240 [214283-224197]	**219205**	239508	
Italy	201505	209946 [206707-213185]	**210717**	232434	
United Kingdom	161145	180122 [177596-182648]	186599	318644	
Germany	157641	162370 [160428-164311]	**163175**	194071	
France	126835	131599 [129385-133812]	**131287**	163290	
Belgium	49227	51399 [49488-53309]	**50990**	67095	
Netherlands	38416	39918 [38416-42206]	**40571**	52600	
Switzerland	29181	29554 [29278-29830]	**29822**	32352	
Portugal	24324	25497 [24324-26855]	**25524**	28376	
Ireland	19877	21232 [19896-22569]	**21506**	39108	

https://doi.org/10.1371/journal.pcbi.1008431.t001

quantity, and, second, the determination of whether the real quantity was within the error of the prediction. These two indexes allow us to calibrate the error bars of the model since we can calculate the percentage of success.

To construct the predictions we used all the data available from the day where cumulative cases crossed the threshold value of 100 cases. However, the successive changes in the control measures could affect the parametrization of the curves. We improved the predictions employing only the last 15 values of the data, after the start of local transmission in the epidemics.

In Fig 6 we show the relative error of the predictions with respect to real data. First, we obtain relative errors for the prediction for the next day of around 2%. The error increases for the predictions for the next days up to the average error of around 5% for the fifth day; see Fig 6A.

The predictions are obtained with a certain error due to uncertainty in the estimation of the parameters of the Gompertz function. Therefore, we evaluated, in Fig 6B, the probability of the actual value being within the prediction intervals around the predicted value. The probability for the first day is around 90% of confidence while this probability decreases for the next days to around 60% for the fifth day; see Fig 6B. We certainly were successful in predictions at short-times of the cumulative cases and therefore the new cases, and, as expected, the accuracy of the predictions decayed over time.

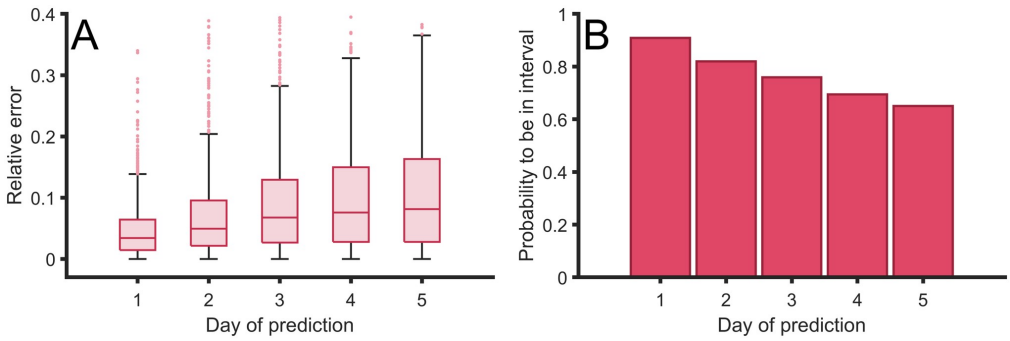

Fig 6. Error of the predictions done by the dynamical fitting of Gompertz function. (A) Relative error between the predictions of the confirmed case for the next five days, in comparison with the actual confirmed cases in several countries. (B) Probability of obtaining the actual real value within the interval of confidance inside the error bars for the next five days. Errors computed with retrospective using all countries with over 1000 cases on April 9, 2020 using ECDC reported cases [28].

https://doi.org/10.1371/journal.pcbi.1008431.g006

Short-term prediction error is corrected with filters

For the predictions made in the previous section on a given day, we used the reported data from 15 days before in order to fit the parameters of the Gompertz function, giving the same weight to all 15 days. From the methodological point of view we improved our predictions using filters to give more relevance to the last data points. We were able to give more weight to the last days and compare with the prediction considering all data points with the same standing. This may be especially useful to rapidly capture changes in trends, as for instance those that we found around the peak of new cases.

We tried several options and concluded that three different filters must be analyzed. We proceeded to show how they behave using the data sets for different countries available. The first filter consists of linear increase in weight between the first and the fifteenth day, the second one a parabolic growth of the weight, and, finally, the third one gives more relevance to only the last three days (a hundred times larger than the other twelve days). By comparison with the equal weight and the other three filters, we obtained a filter which minimizes the relative error; see the comparison in Fig 7A.

Although comparison among the four procedures, see Fig 7, shows relatively small differences, this statistical study shows performs better with the last filter, which gives greater weight to the last three values of the data. The performance of such filter is particularly better when the epidemic approaches the values of the peak of new cases.

The average of the relative error decreases with the asymmetry of the type of filter we employ; see Fig 7A. The filter with greatest weights in the last three events presents a better performance in comparison with the other three filters employed. It is important to note that we also checked other filters with greater weight in the last single event and the last two events, and the results were less accurate.

We obtain similar results if we evaluate the probability of success of the predictions of each filter; see Fig 7B. Light bars in such a figure show success using the error bars obtained from the mean square method adapted to each of the filters. Note that the error bars, or confidence intervals, of each method may be different and therefore this may affect the likelihood of success because it produces larger confidence intervals. To systematically compare the four methods we employed the confidence interval of the original method with the mean values

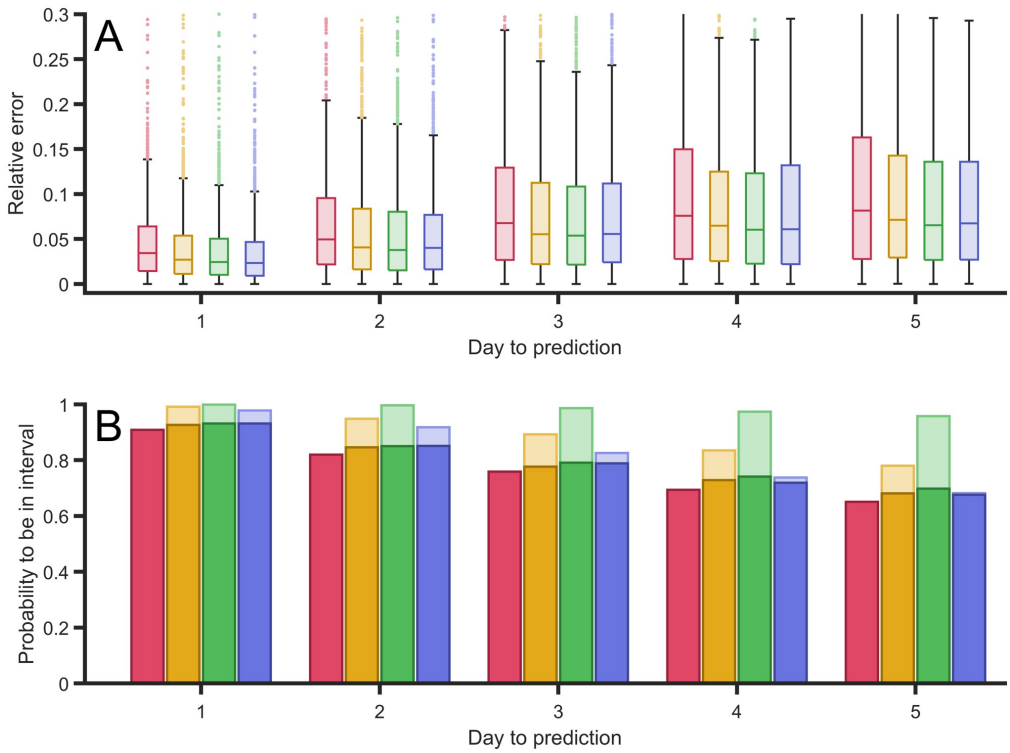

Fig 7. Error of the predictions with different type of filters. (A) Relative error between the predictions of the confirmed case for the next five days, in comparison with the actual confirmed cases for four different types of filters: constant 15 values (red), linear increase (orange), parabolic increase (green), and a filter with three largest last values (blue). (B) Probability of obtaining the actual real value within the interval of confidance inside the error bars for the next five days using the same four filters. Light bars to the probability of being found within the confidence interval using each filter confidance interval;, dark bars show probability of being within the confidence interval using first filter confidance interval to be able to compare among the different filters. Although different filters have different confidence intervalsizes, they have the same significance level of $\alpha = 0.01$.

https://doi.org/10.1371/journal.pcbi.1008431.g007

obtained in the other filters. Note that with such definition, the dark and light bars for the first method overlap. We also observed better performance in increasing the asymmetry of the filter and as in the previous comparison, the method focused on the last three values maximized the probability of success.

Discussion

Finally, we discuss the possibility of longer-term estimations with the Gompertz function, and offer our main conclusions.

Long-term estimations can be obtained from Gompertz model

We are assuming simple premises and they permit us to expand the short-term predictions shown above and calculate longer-term predictions greater than five days, explicitly the values of K, t_p and 90%K. Long-term estimations are possible during a certain outbreak wave. Second

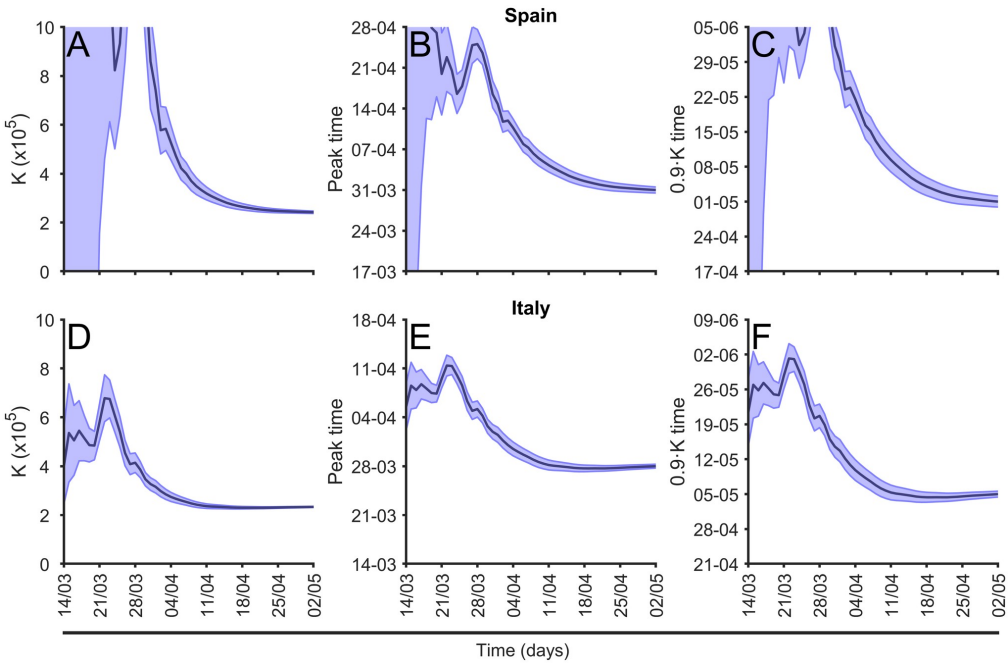

Fig 8. Evolution of the long-term estimations. (A) and (D) Evolution of the prediction of final total number of cases, K; (B) and (E) evolution of prediction of the time for the peak of maximum new cases prediction, see Eq (6) and (C) and (F) evolution of the arrival to 90% of total cases between March 14, 2020 and May 2, 2020 in Spain and Italy, respectively.

https://doi.org/10.1371/journal.pcbi.1008431.g008

waves may completely change the dynamics and the values of final incidence. Such new epidemic focuses are not considered in the model, we may treat them as an independent epidemic for which the numbers probably have to be reset.

The use of a phenomenological function facilitates the projection to the future of the trend in comparison with other methods which evaluate in the vicinity of the last day. Although the only relatively reliable predictions in such a complicated problem are short-term predictions, we can however address relevant questions like the final value of total cases of parameter K, predictions of the peak or maximum of new cases, or the time needed to arrive at 90% of the total cases. To obtain such long-term estimations we employ the whole data set for each country to unveil the trend of the whole dynamics.

We calculate daily the parameter values of the fitting function described above and the evolution of the parameter K for different countries together with two characteristic times of the epidemic. See two examples, Spain and Italy, in Fig 8, for the value of K, t_p, and 90%K. For other countries in Europe see, respectively, supplementary S3 Fig, supplementary S4 Fig, and supplementary S5 Fig. The estimations begin with large uncertainty; however, the values converge on the actual value systematically for the three calculations. The confidence interval also reduces with time, although there are systematic fonts of errors not addressed by the interval. The main differences between Spain and Italy in Fig 8 are the large errorbars for Spain at the beginning of the evolution, because of the delay in the epidemic phase of both countries on

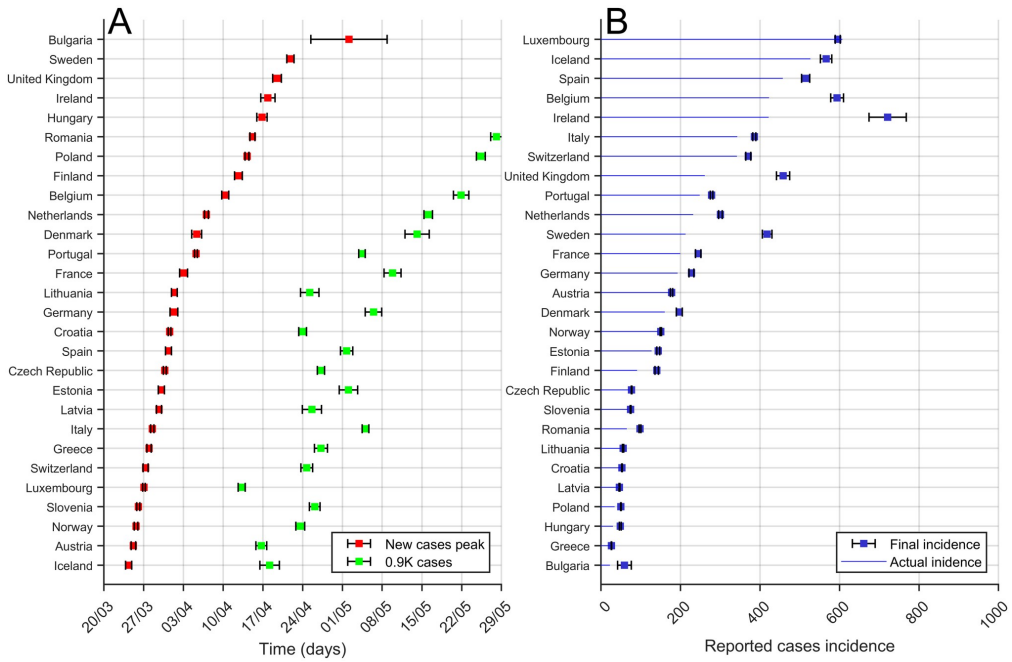

Fig 9. Comparison of long-term estimations among European Countries. (A) Time for the peak prediction (red) and the time to arrive at 90% of total cases (in green) predictions obtained from the last evaluation of the Gompertz function (April 12, 2020) to the evolution of the cumulative cases. Countries are sorted from top to bottom using time-to–peak-time prediction. (B) Final incidence (total cases per 10^5 inhabitants) prediction (blue squares) obtained from the last evaluation of the Gompertz function (April 12, 2020) to the evolution of the cumulative cases (blue line); see procedure in Fig 8. Error bars correspond to the error obtained from the fit and the corresponding error propagation. Countries are sorted from top to bottom in terms of actual incidence.

https://doi.org/10.1371/journal.pcbi.1008431.g009

March 9th, when the graphic begins. While in Italy the epidemic was fully developed, in Spain the epidemic was at the initial phase with an exponential growth.

Using the method described above we can compare the three predictions shown in Fig 8 for all the countries in Europe for a particular date; see this comparison in Fig 9. For the two temporal comparisons note that actually the dates for the peak for some countries had already been passed at the time when the evaluation was made. However, it is actually not always clear when the actual moment a country is passing the peak is. Furthermore, for the comparison among the different countries in Europe with very different demographics, we used the incidence of the epidemic, evaluated as the number of cases per 10^5 inhabitants. In this graphic we compare with the actual phase of the epidemic at May 2, 2020 in each country [22]. While some of the countries are close to the final number of cases, there are some countries still at the initial phase of the epidemic with very large growth, which predicts large incidence rates. This is the case with the United Kingdom.

Note that we have to approach the previous estimations reticently, because they are only approximations assuming some simple premises. Therefore, we consider such estimations as objects for discussion rather than as results of the model.

Outlook and conclusions

We fitted the Gompertz function to the cumulative cases in different regions and countries to be able to infer, from the fitted parameters of the model, relevant quantities for the understanding of the epidemics. On the one hand, we obtained reliable short-time predictions for the new cases during the subsequent days. These predictions are robust and the percentage of success is around 90% for the next day. but in addition, the fitting provides some long-term quantities, for example, estimations of the total number of cases or the timing of the peak of new cases.

As an empirical function, Gompertz does not depend on previous knowledge of the system. It is especially useful in situations where there is no deep knowledge of the internal structure of the epidemics and when key properties of the epidemics are not known. It is precisely the lack of knowledge regarding the different pathways of contagion or its dependence on social measure that makes the fitting of a quantitative predictable model impossible. Complex models with a lot of parameters to fit are, in this type of epidemic, exercises in exploring possible scenarios, but never real quantitative tools. No model can predict the reaction of the population to a particular measure, nor even properly assess the parameters of mobility when even basic immunity questions remain unsolved. This is what makes our results about the large degree of confidence in terms of short-term predictions of the evolution of the Covid-19 epidemics so important. Our work has important ramifications since it can predict, and at the same time assess, changes in the dynamics of the pandemic. The prediction procedure adapts to changes in any of the structural properties of the system. Changes in the diagnostic testing needed to detect a case, in social measures, or in the way of counting cases just introduce variation in the model that fades away as the new properties emerge again. We have clearly shown in this paper that this changing structure is properly captured with the decreasing nature of the growth given by parameters μ_0 and a, and the final number of cases K. The highly complex and unknown nature of key elements of the epidemics does not prevent our predicting its evolution in the short-term and to assess the control, or lack thereof, of the epidemics' spread.

We may conclude that the methodology here presented can be further employed for the evaluation of the epidemic and the control measures in the next countries to which it spreads in its initial stage. We obtain predictions with a success greater than 90%, which means that around 90% of the reported cases are within the prediction intervals.

We are planning to further collaborate with health institutions in Africa and America to advise them with the predictions of the model for the evolution of the Covid-19 epidemic in these countries. In such collaboration, the continuous interplay between predictions and results during spreading will lead us to a rethinking of the assumptions of our model. We hope to further improve the predictions by the introduction of changes, if needed. Further work can be done to improve the prediction process. The results of the fitting might be better if country-wide data is disaggregated for more homogeneous subnational regions. Data shows that in some countries the appearance of different focuses produces the formation of different epidemics which under the conditions of strong restriction of movement can give rise to independent dynamics within the country. It is more reliable to work with information at the regional level although the number of cases is lower and the fluctuations stronger. We have observed good statistical behavior and predictions in the secondary outbreak in several Catalan cities [36] with a total population of around half a million or more. We expect that models applied to regions of this size could be useful to predict more aggregated scales. However, the main limitation of the regional approach up to now has been the lack of detailed data and/or the difference in the protocols and definitions used by local authorities. As the pandemic advanced more reliable data at the regional level was available; see for example [37].

Finally, we would like to note that the use a generic function is an empirical tool to treat future local and global epidemics, as has been begun recently with other growth functions like the Verhust and Richards models [17]. We plan to continuously update the approach employed here to adapt to any special particularity of any new epidemics. Presently, the same data are applied to guide public policy in hospital administrations giving assessment to regional governments regarding the short-term evolution of health needs.

In order to take adequate and precise control measures political leaders need up-to-date information on the epidemics and a clear representation of the phase of the epidemics among several countries or in a particular country of different regions. We have found our short-time predictions to be a highly valuable information tool for policymakers, since they it can help guide their short-term planning decisions.

Supporting information

S1 Fig. Cases in different European countries. The total cases together with the new daily cases with the corresponding fitings obtained from the Gompertz model are shown for a selection of European countries.
(TIF)

S2 Fig. Evolution of the fitting of parameter *a*. The dynamics of the fitting of parameter a obtained from fitting from the Gompertz model are shown for a selection of European countries.
(TIF)

S3 Fig. Evolution of the fitting of parameter *K*. The dynamics of the fitting of parameter K obtained from fitting from the Gompertz model are shown for a selection of European countries.
(TIF)

S4 Fig. Evolution of the fitting of parameter t_p. The dynamics of the fitting of parameter t_p obtained from fitting from the Gompertz model are shown for a selection of European countries.
(TIF)

S5 Fig. Evolution of the fitting of parameter 90%*K*. The dynamics of the fitting of the parameter 90%K obtained from fitting from the Gompertz model are shown for a selection of European countries.
(TIF)

Author Contributions

Conceptualization: Martí Català, Sergio Alonso, Enrique Alvarez-Lacalle, Daniel López, Pere-Joan Cardona.

Formal analysis: Martí Català, Sergio Alonso, Enrique Alvarez-Lacalle, Daniel López.

Funding acquisition: Martí Català, Sergio Alonso, Daniel López, Pere-Joan Cardona, Clara Prats.

Investigation: Martí Català, Sergio Alonso, Enrique Alvarez-Lacalle, Daniel López, Pere-Joan Cardona, Clara Prats.

Methodology: Martí Català, Sergio Alonso, Enrique Alvarez-Lacalle, Daniel López, Pere-Joan Cardona, Clara Prats.

Software: Martí Català, Clara Prats.

Writing – review & editing: Martí Català, Sergio Alonso, Enrique Alvarez-Lacalle, Daniel López, Pere-Joan Cardona, Clara Prats.

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020; 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5
2. Doms C, Kramer SC, Shaman J. Assessing the use of influenza forecasts and epidemiological modeling in public health decision making in the United States. Scientific reports. 2018; 8(1):1–7. https://doi.org/10.1038/s41598-018-30378-w
3. Anderson RM, Anderson B, May RM. Infectious diseases of humans: dynamics and control. Oxford university press; 1992.
4. Lekone PE, Finkenstädt BF. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics. 2006; 62(4):1170–1177. https://doi.org/10.1111/j.1541-0420.2006.00609.x
5. Althaus CL. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. PLoS currents. 2014; 6.
6. Ng TW, Turinici G, Danchin A. A double epidemic model for the SARS propagation. BMC Infectious Diseases. 2003; 3(1):19. https://doi.org/10.1186/1471-2334-3-19
7. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science. 2003; 300 (5627):1961–1966. https://doi.org/10.1126/science.1086478 PMID: 12766206
8. Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003; 300(5627):1966–1970. https://doi.org/10.1126/science.1086616 PMID: 12766207
9. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling. 2020; 5:256–263. https://doi.org/10.1016/j.idm.2020.02.002 PMID: 32110742
10. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. Plos one. 2020; 15(3): e0231236. https://doi.org/10.1371/journal.pone.0231236
11. Català M, Pino D, Marchena M, Palacios P, Urdiales T, Cardona PJ, Alonso S, Lopez-Codina D, Prats C, Alverz-Lacalle E. Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. 2020;. medRxiv
12. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PloS one. 2020; 15(3):e0230405. https://doi.org/10.1371/journal.pone.0230405
13. Verhulst PF. Notice sur la loi que la population suit dans son accroissement. Corresp Math Phys. 1838; 10:113–126.
14. Viboud C, Simonsen L, Chowell G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics. 2016; 15:27–37. https://doi.org/10.1016/j.epidem.2016.01.002
15. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling. 2017; 2(3):379–398. https://doi.org/10.1016/j.idm.2017.08.001
16. Wang XS, Wu J, Yang Y. Richards model revisited: Validation by and application to infection dynamics. Journal of Theoretical Biology. 2012; 313:12–19. https://doi.org/10.1016/j.jtbi.2012.07.024
17. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020. Journal of Clinical Medicine. 2020; 9(2):596. https://doi.org/10.3390/jcm9020596 PMID: 32098289
18. Gompertz B. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. Philosophical transactions of the Royal Society of London. 1825;(115):513–583.
19. Bürger R, Chowell G, Lara-Díaz LY. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. Mathematical biosciences and engineering: MBE. 2019; 16(5):4250–4273. https://doi.org/10.3934/mbe.2019212

20. Liu W, Tang S, Xiao Y. Model selection and evaluation based on emerging infectious disease data sets including A/H1N1 and Ebola. Computational and mathematical methods in medicine. 2015; 2015. https://doi.org/10.1155/2015/207105

21. Pell B, Kuang Y, Viboud C, Chowell G. Using phenomenological models for forecasting the 2015 Ebola challenge. Epidemics. 2018; 22:62–70. https://doi.org/10.1016/j.epidem.2016.11.002

22. Prats C, Alonso S, Álvarez-Lacalle E, Marchena M, López-Codina D, Català M, Cardona PJ, Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries. Universitat Politècnica de Catalunya; April 22 2020.; 2020. Available from: https://upcommons.upc.edu/handle/2117/110978.

23. Levitt M, Scaiewicz A, Zonta F Predicting the trajectory of any COVID19 epidemicfrom the best straight line medRxiv

24. Ohnishi A, Namekawa Y, and Fukui T. Universality in COVID-19 spread in view of the Gompertz function medRxiv

25. Torrealba-Rodriguez, O., Conde-Gutiérrez R A, and Hernández-Javier A L. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models Chaos, Solitons and Fractals (2020): 109946.

26. Dutra, C M, Farias F M, and Riella de Melo C A. New approach of non-linear fitting to estimate the temporal trajectory of the COVID-19 cases Brazilian Journal of Health Review 3.3 (2020): 6341-6356.

27. Data on the geographic distribution of COVID-19 cases worldwide Available from: https://github.com/catalamarti/Gompertz_Catala2020/blob/main/Data_ECDC.xlsx.

28. Download today's data on the geographic distribution of COVID-19 cases worldwide European Centre for Disease Prevention and Control.; 2020. Available from: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

29. Coronavirus disease (COVID-2019) situation reports. Available from: https://covid19.who.int/table.

30. Zwietering M, Jongenburger I, Rombouts F, Van't Riet K. Modeling of the bacterial growth curve. Appl Environ Microbiol. 1990; 56(6):1875–1881. https://doi.org/10.1128/AEM.56.6.1875-1881.1990

31. Gerlee P. The model muddle: in search of tumor growth laws. Cancer research. 2013; 73(8):2407–2411. https://doi.org/10.1158/0008-5472.CAN-12-4355

32. Taylor J. An Introduction to Error Analysis: The Study of Uncertainties in; 1982.

33. Madden L, et al. Quantification of disease progression. Protection Ecology. 1980; 2(n, 1).

34. Prats C, Alonso S, López-Codina D, Català M, Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries. 1. Universitat Politècnica de Catalunya; April 22 2020.; 2020. Available from: https://upcommons.upc.edu/handle/2117/186486.

35. Prats C, Alonso S, Álvarez-Lacalle E, Marchena M, López-Codina D, Català M, Cardona PJ, Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries. 16. Universitat Politècnica de Catalunya; April 22 2020.; 2020. Available from: https://upcommons.upc.edu/handle/2117/186488.

36. Prats C, Alonso S, Álvarez-Lacalle E, Marchena M, López-Codina D, Català M, Conesa D, Cardona PJ, Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries. 103. Universitat Politècnica de Catalunya; July 17 2020.; 2020. Available from: https://upcommons.upc.edu/handle/2117/327043.

37. Joint Research Center ECML Covid; 2020 Regional map available from: https://webcritech.jrc.ec.europa.eu/modellingoutput/cv/eu_cv_region/eu_cv_region_inf.htm.

Part III.
Conclusions

11 Conclusions and outcomes

In this thesis several methods for understanding, predicting, and conditioning the dynamics of TB and COVID-19 are proposed. This thesis covers a wide range of different models to tackle the different problems that we faced or the gaps that we need to fill.

The specific conclusions of this thesis are:

1. **Immune response balance is crucial to determine the outcome of a tuberculosis infection.**
   We have built an alveolus level reaction model that determines three possible outcomes: proliferative lesion (*Mtb* are eliminated), exudative lesion (no control), and bistability (both states are possible). The parameters that determine the final outcome were identified. These parameters are mainly related with the immune response.

2. **Encapsulation process lead by fibroblasts can contain the tuberculosis lesions growth, while the distance to the membranes where fibroblasts are located conditions lesion size.**
   We have built a secondary lobe reaction-diffusion model that includes the encapsulation process. Encapsulation leaves a non-growing lesion formed by different cells covered by fibroblasts. Size of encapsulated lesions mainly depends on two factors: (1) the distance to the nearest pulmonary membrane (where fibroblasts are) and (2) diffusion coefficients of the different cells that form the lesion. Some immune response parameters also help to reduce lesion size.

3. **Dynamic hypothesis is confirmed as a possible explanation for the temporal evolution of lesions.**
   A computational hybrid model has been built to simulate the development of tuberculosis lesions in a latent tuberculosis infection in minipigs. Simulated data reproduced the experimental measured lesion sizes and positions. We tested the dynamic hypothesis falsifiability and confirmed it is a possible explanation for the dynamics of TB infection in lungs. We have not been able to determine criteria to distinguish the lesions that were part of the initial infection.

4. **An exaggerated inflammatory response or a moderate inflammatory response combined with a small breathing amplitude can cause the active tuberculosis disease.**
   Transition between latent tuberculosis infection (LTBI) and ATB has been explored and the parameters that caused ATB have been identified. According to our model there are two parameters that can cause an ATB: (1) an exaggerated inflammatory response or (2) a moderated inflammatory response combined with an small breathing amplitude, which increases the probability that lesions merge and bigger lesions are formed.

5. **Female resistance and protection are key points to assure the coexistence of *Mtb* and humans.**
   Continuous and discrete models have been built to simulate the co-existence of humans and tuberculosis. Female resistance and protection were identified as key points to assure this coexistence. In the paleolithic age, if this factor did not exist tuberculosis would have extinguished the communities where it appeared. In fact, it has also been seen that the emergence of new, more contagious lineages in the paleolithic was not possible due to the small birth rate.

6. **"Modern" lineages of *Mtb* quickly overtake the "ancient" ones, but their appearance in the paleolithic is not possible.**
   A competition model between different tuberculosis lineages (treated as independent diseases) was built. The substitution of "ancient" *Mtb* lineages by the so-called "modern" lineages was not possible during the paleolithic age as seen in the previous model. In this new model we have seen that "modern" lineage quickly imposed their hegemony over the "ancient" one that was relegated but not eliminated.

7. **It is estimated that 17 million Europeans had been infected from COVID-19 before June 2020.**
   Total COVID-19 incidence has been estimated from reported COVID-19 mortality in most countries in Europe. It is estimated that 17 million Europeans ($\sim$2.2% of total population) had COVID-19 before June 2020. Some countries had greater incidence; for example, in Spain it is estimated that 6.2% of the population had COVID-19 before June 2020, and in Italy around 5.6%. Seroprevalence studies have confirmed these numbers.

8. **The COVID-19 diagnostic rates of European countries during the first wave were successfully determined.**
   Delays between case, detection and death have been identified for each of the studied countries. During this called first wave just 8% of the cases were diagnosed. This number had a wide range between different countries, from the 4% of cases detected in France or United Kingdom to 25% in Germany. On 20 April 2020 there were estimated a total of 4.5 million active cases (cumulative cases prior 14 days) in 10 countries in Europe, most of them in the United Kingdom.

9. **Gompertz model fits the evolution of COVID-19 pandemic.**
   We correctly fitted the evolution of the COVID-19 epidemic in China and its different provinces with the Gompertz model. We computed R^2 in each of the Chinese provinces to measure the goodness of fit. We saw that R^2 for the cumulative case count was bigger than 0.99 in each territory and for the new case count we saw that values were around 0.9. This empirical model has been adjusted to the new growing epidemics in European countries, but we have seen that uncertainty of trial parameters is too high to make a precise prediction of how many cases are going to be measured during the observed first wave.

10. **Gompertz predictions had less than 8% of relative error on the subsequent 5-day predictions.**
    Extending the fitting to the future we have been able to make predictions. A retrospective study to compute the prediction error for all world countries that reported more than 1,000 cases on 9 April 2020 was carried out. Expected prediction error values were determined, for example it was determined that with this model we had less than 8% relative error in the cumulative case count for 5 days in the future.

12 Perspectives

Mathematical models were used to understand different stages in the evolution of tuberculosis, from the lowest alveolus level to the epidemiological level. We were able to examine a biological hypothesis using computational models for its falsifiability. It is expected that these results can help to reduce the use of animal models and avoid potential harm to the animals. Using mathematical models, the experiments can be refined and gain in quality and utility.

These analyses were carried out in the Comparative Medicine and Bioimage Centre of Catalonia (CMCiB). This thesis is an example of the usefulness of computational models. In the near future the CMCiB plans to use these types of models intensively to solve other research questions and to continue on the path of the 3Rs (Reduction, Refinement and Replacement) in animal models. These new projects where the 3R framework will be employed are in diverse current research fields in the CMCiB such as ictus and genomics. We also plan the use of these models in the research on tuberculosis; we aim to refine the current models to better reproduce tuberculosis infection inside the lungs and to analyze the processes involved.

The COVID-19 models shown in this thesis were used from March 2020 to present in order to forecast pandemic evolution and estimate the real incidence and diagnostic rate in many countries. They were also refined and are in continuous evolution to be adapted to new pandemic situations.

Despite all the efforts made by the governments and their population, the COVID-19 pandemic has not ended yet. During the coming months the BIOCOM-SC group will continue with the preparation of reports for the DG-CONNECT European Commission department and to the health institutions in the country. The models developed in this thesis will be useful in preparing these reports and predicting COVID-19 pandemic evolution.

This thesis shows that mathematical and computational models are useful in biomedicine and epidemiology. We have been able to address different problems and questions in these fields. From the understanding of the dynamics of tuberculosis infection in an alveolus to forecasting the evolution of COVID-19 pandemic in the following days. We hope that

the results of this thesis highlight the importance of computational and mathematical models and to help advance their use in many different areas of research.

Bibliography

[1] Salim S Abdool Karim i Tulio de Oliveira. "New SARS-CoV-2 variants—clinical, public health, and vaccine implications". A: *New England Journal of Medicine* (2021).

[2] Agència de Qualitat i Avaluació Sanitàries de Catalunya. *Evolució dels casos i de la Rt del SARS-CoV2*. `https://app.powerbi.com/view?r=eyJrIjoiZTE2OTAwYzctMzQ3ZC00ZjY5LWExNDctYWIzODFmMDQxMjRmIiwidCI6IjNiOTQyN2RjLWQzMGUtNDNiYy04YzA2LWZmNzI1MzY3NmZlYyIsImMiOjh9`. Online; accessed 13 May 2021. 2021.

[3] Nura MR Ahmad et al. "Analyzing policymaking for tuberculosis control in Nigeria". A: *Complexity* 2018 (2018).

[4] Amani M Alnimr. "Dormancy models for Mycobacterium tuberculosis: A minireview". A: *Brazilian Journal of Microbiology* 46.3 (2015), pàg. 641 - 647.

[5] Christian L Althaus. "Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa". A: *PLoS currents* 6 (2014).

[6] Allisson Dantas. *BIOCOMSC-COVID19-MAT2PY*. `https://github.com/allissondantas/BIOCOMSC-COVID19-MAT2PY`. Online; created 21 December 2020. 2020.

[7] Roy M Anderson et al. "Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic". A: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359.1447 (2004), pàg. 1091 - 1105.

[8] Lindsey R Baden et al. "Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine". A: *New England Journal of Medicine* 384.5 (2021), pàg. 403 - 416.

[9] V Balasubramanian et al. "Pathogenesis of tuberculosis: pathway to apical localization". A: *Tubercle and Lung Disease* 75.3 (1994), pàg. 168 - 178.

[10] Anirban Basu. "Estimating The Infection Fatality Rate Among Symptomatic COVID-19 Cases In The United States: Study estimates the COVID-19 infection fatality rate at the US county level." A: *Health Affairs* 39.7 (2020), pàg. 1229 - 1236.

[11] Matthew Bates, Ben J Marais i Alimuddin Zumla. "Tuberculosis comorbidity with communicable and noncommunicable diseases". A: *Cold Spring Harbor perspectives in medicine* 5.11 (2015), a017889.

[12] Jordi Bechini. "Estudio de la Tuberculosis pulmonar mediante Tomografía Computarizada Multidetector en un modelo experimental de minipig". Tesi doct. Universitat Autònoma de Barcelona, 2016.

[13] John H Beigel et al. "Remdesivir for the treatment of Covid-19". A: *New England Journal of Medicine* 383.19 (2020), pàg. 1813 - 1826.

[14] George I. Bell. "Cell Growth and Division: III. Conditions for Balanced Exponential Growth in a Mathematical Model". A: *Biophysical Journal* 8.4 (1968), pàg. 431 - 444. ISSN: 0006-3495. DOI: https://doi.org/10.1016/S0006-3495(68)86498-7. URL: https://www.sciencedirect.com/science/article/pii/S0006349568864987.

[15] Shmuel Benenson et al. "BNT162b2 mRNA Covid-19 vaccine effectiveness among health care workers". A: *New England Journal of Medicine* (2021).

[16] Luiz E Bermudez, Lia Danelishvili i Julie Early. "Mycobacteria and macrophage apoptosis: complex struggle for survival". A: *Microbe (Washington, DC)* 1.8 (2006), pàg. 372.

[17] Ashu Seith Bhalla et al. "Chest tuberculosis: Radiological review and imaging recommendations". A: *The Indian journal of radiology & imaging* 25.3 (2015), pàg. 213.

[18] Alessandro Boianelli et al. "Modeling influenza virus infection: a roadmap for influenza research". A: *Viruses* 7.10 (2015), 5274 - 5304.

[19] Vítor Borges et al. "Tracking SARS-CoV-2 VOC 202012/01 (lineage B. 1.1. 7) dissemination in Portugal: insights from nationwide RT-PCR Spike gene drop out data". A: *Euro. Surveill* 26 (2021), pàg. 2100131.

[20] Daniela Brites i Sebastien Gagneux. "Co-evolution of M ycobacterium tuberculosis and H omo sapiens". A: *Immunological reviews* 264.1 (2015), pàg. 6 - 24.

[21] Roland Brosch et al. "A new evolutionary scenario for the Mycobacterium tuberculosis complex". A: *Proceedings of the national academy of Sciences* 99.6 (2002), pàg. 3684 - 3689.

[22] Antonio Bru i Pere-Joan Cardona. "Mathematical modeling of tuberculosis bacillary counts and cellular populations in the organs of infected mice". A: *PLoS One* 5.9 (2010), e12985.

[23] John F Brundage i G Dennis Shanks. "Deaths from bacterial pneumonia during 1918–19 influenza pandemic". A: *Emerging infectious diseases* 14.8 (2008), pàg. 1193.

[24] Raimund Bürger, Gerardo Chowell i Leidy Yissedt Lara-Díiaz. "Comparative analysis of phenomenological growth models applied to epidemic outbreaks." A: *Mathematical biosciences and engineering: MBE* 16.5 (2019), pàg. 4250 - 4273.

[25] Carmen Cabezas et al. "Effects of BNT162b2 mRNA Vaccination on COVID-19 Disease, Hospitalisation and Mortality in Nursing Homes and Healthcare Workers: A Prospective Cohort Study Including 28,594 Nursing Home Residents, 26,238 Nursing Home Staff, and 61,951 Healthcare Workers in Catalonia". A: *Available at SSRN 3815682 (pre-print)* (2021). DOI: http://dx.doi.org/10.2139/ssrn.3815682.

[26] Peter P Calow. *Encyclopedia of ecology and environmental management*. John Wiley & Sons, 2009.

[27] Ewen Callaway i Smriti Mallapaty. "NOVAVAX COVID VACCINE PROTECTS PEOPLE AGAINST VARIANTS". A: *Nature* 590 (2021), pàg. 17.

[28] Pere-Joan Cardona. "Revisiting the natural history of tuberculosis". A: *Archivum immunologiae et therapiae experimentalis* 58.1 (2010), pàg. 7 - 14.

[29] Pere-Joan Cardona i Clara Prats. "The small breathing amplitude at the upper lobes favors the attraction of polymorphonuclear neutrophils to Mycobacterium tuberculosis lesions and helps to understand the evolution toward active disease in an individual-based model". A: *Frontiers in microbiology* 7 (2016), pàg. 354.

[30] Pere-Joan Cardona et al. "Can systems immunology lead tuberculosis eradication?" A: *Current opinion in systems biology* 12 (2018), pàg. 53 - 60.

[31] Carlos III Health Institute. *Número de hospitalizaciones, número de ingresos en UCI y número de defunciones por sexo, edad y provincia de residencia*. https://cnecovid.isciii.es/covid19/resources/casos_hosp_uci_def_sexo_edad_provres.csv. Online; accessed 13 May 2021. 2021.

[32] Boye Bobby Carter et al. “Survival analysis of patients with tuberculosis and risk factors for multidrug-resistant tuberculosis in Monrovia, Liberia”. A: *Plos one* 16.4 (2021), e0249474.

[33] Marti Catala et al. “The Impact of Prioritisation and Dosing Intervals on the Effects of COVID-19 Vaccination in Europe: An Agent-Based Cohort Model”. A: *Available at SSRN 3793540 (preprint)* (2021). DOI: http://dx.doi.org/10.2139/ssrn.3793540.

[34] Martí Català. “A 3D computational model for understanding tuberculosis lesions dynamics in lungs”. Bachelor’s Thesis. Universitat de Barcelona, 2016.

[35] Martí Català. “Machine Learning per a l’optimització d’un model epidemiològic en tuberculosi”. Bachelor’s Thesis. Universitat Oberta de Catalunya, 2018.

[36] Martí Català. “Modelling and simulation of tuberculosis lesions dynamics in a minipig bronchial tree”. Bachelor’s Thesis. Universitat Politècnica de Catalunya, 2015.

[37] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 141*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/330273.

[38] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 141*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/186486.

[39] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 152,On the weekend effect on confirmed cases and the resulting oscillations in the empiric reproduction number*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/346361.

[40] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 154, On the weekend effect on confirmed cases and the resulting oscillations in the empiric reproduction number (II)*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/346361.

[41] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 155, On the weekend effect on confirmed cases. Part III: Rethinking short-term prediction methodology to account for this effect*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/346361.

[42] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 230*. Inf. tèc. Universitat Politècnica de Catalunya, 2021. URL: http://hdl.handle.net/2117/345659.

[43] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 38*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/184991.

[44] Martí Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 42, Weekend effect and data analysis in an epidemic*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: http://hdl.handle.net/2117/346361.

[45] Catalan Health Institute. *Salut/Dades COVID, Registre de casos de COVID-19 realitzats a Catalunya*. https://dadescovid.cat/descarregues. Online; accessed 13 May 2021. 2021.

[46] Center of Disease Control and Prevention. *National Center for Immunization and Respiratory Diseases (NCIRD)*. https://www.cdc.gov/ncird/. Online; accessed 13 May 2021. 2021.

[47] Center of Disease Control and Prevention. *National Center for Immunization and Respiratory Diseases (NCIRD)*. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/recommendations-process.html. Online; accessed 13 May 2021. 2021.

[48] Central Intelligence Agency. *The world factbook, County comparisons - Median age*. https://www.cia.gov/the-world-factbook/field/median-age/country-comparison. Online; accessed 13 May 2021. 2021.

[49] Central Intelligence Agency. *The world factbook, County comparisons - Real GDP per capita*. https://www.cia.gov/the-world-factbook/field/real-gdp-per-capita/country-comparison. Online; accessed 13 May 2021. 2021.

[50] Iñaki Comas et al. “Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans”. A: *Nature genetics* 45.10 (2013), pàg. 1176 - 1182.

[51] Corporació Catalana de Mitjans Audiovisuals. *Els casos de Covid-19 diagnosticats a Catalunya serien només un 10% dels reals*. https://www.ccma.cat/324/els-casos-de-covid-19-diagnosticats-a-catalunya-serien-nomes-un-10-dels-reals/noticia/3008205/. Online; published 29 April 2020. 2020.

[52] Alex Crozier et al. “Put to the test: use of rapid testing technologies for covid-19”. A: *bmj* 372 (2021).

[53] David M Cutler i Lawrence H Summers. “The COVID-19 pandemic and the $16 trillion virus”. A: *Jama* 324.15 (2020), pàg. 1495 - 1496.

[54] M. H. Cynamon et al. “Activities of Several Novel Oxazolidinones againstMycobacterium tuberculosis in a Murine Model”. A: *Antimicrobial Agents and Chemotherapy* 43.5 (1999), pàg. 1189 - 1191. ISSN: 0066-4804. DOI: 10.1128/AAC.43.5.1189. eprint: https://aac.asm.org/content/43/5/1189.full.pdf. URL: https://aac.asm.org/content/43/5/1189.

[55] Po-Lin Chen et al. “A Review of Treatment of Coronavirus Disease 2019 (COVID-19): Therapeutic Repurposing and Unmet Clinical Needs”. A: *Frontiers in pharmacology* 11 (2020).

[56] Nanshan Chen et al. “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study”. A: *The lancet* 395.10223 (2020), pàg. 507 - 513.

[57] Jee Young Chung, Melissa N Thone i Young Jik Kwon. “COVID-19 vaccines: The status and perspectives in delivery points of view”. A: *Advanced drug delivery reviews* (2020).

[58] Danish Covid-19 Genome Consortium. *Genomic overview of SARS-CoV-2 in Denmark*. https://www.covid19genomics.dk/statistics. Online; accessed 1 June 2021. 2021.

[59] diari Ara. *Els brots situen Espanya entre els països més afectats d’Europa*. https://www.ara.cat/internacional/brots-covid-19-coronavirus-espanya-paisos-afectats-europa_1_1099414.html. Online; published 20 July 2020. 2020.

[60] Richard C Dicker et al. “Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics”. A: (2006).

[61] Susan E Dorman et al. “Four-Month Rifapentine Regimens with or without Moxifloxacin for Tuberculosis”. A: *New England Journal of Medicine* 384.18 (2021), pàg. 1705 - 1718.

[62] Andrea Freyer Dugas et al. “Influenza forecasting with Google flu trends”. A: *PloS one* 8.2 (2013), e56176.

[63] Epiforecasts team (Centre for Mathematical Modelling of Infections Diseases,mLondon School of Hygiene & Tropcial Medicine) and European Centre for Disease Control and Prevention. *European Covid-19 Forecast Hub*. `https://covid19forecasthub.eu/`. Online; accessed 27 May 2021. 2021.

[64] European Centre for Disease Prevention and Control. *COVID-19 situation update worldwide*. `https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases`. Online; accessed 13 May 2021. 2021.

[65] European Centre for Disease Prevention and Control. *Data on country response measures to COVID-19*. `https://www.ecdc.europa.eu/sites/default/files/documents/response_graphs_data_2021-05-13.csv`. Online; accessed 13 May 2021. 2021.

[66] Anthony Fauci i Allison Inserro. “Challenges and similarities in HIV, COVID-19 crises: a Q&A with Anthony Fauci, MD.” A: *The American Journal of Managed Care* 26.7 (2020), pàg. 282 - 283.

[67] Wilfredo F Garcia-Beltran et al. “Circulating SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity”. A: *MedRxiv* (2021).

[68] Catherine Gebhard et al. “Impact of sex and gender on COVID-19 outcomes in Europe”. A: *Biology of sex differences* 11 (2020), pàg. 1 - 13.

[69] Evangelos J Giamarellos-Bourboulis et al. “Activate: randomized clinical trial of BCG vaccination against infection in the elderly”. A: *Cell* 183.2 (2020), pàg. 315 - 323.

[70] Olga Gil et al. “Granuloma encapsulation is a key factor for containing tuberculosis infection in minipigs”. A: *PloS one* 5.4 (2010), e10030.

[71] Yair Goldberg et al. “Protection of previous SARS-CoV-2 infection is similar to that of BNT162b2 vaccine protection: A three-month nationwide experience from Israel”. A: *medRxiv* (2021).

[72] Benjamin Gompertz. "XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c". A: *Philosophical transactions of the Royal Society of London* 115 (1825), pàg. 513 - 583.

[73] Chaitra Gopalappa et al. "Progression and transmission of HIV/AIDS (PATH 2.0) a new, agent-based model to estimate HIV transmissions in the United States". A: *Medical Decision Making* 37.2 (2017), pàg. 224 - 233.

[74] Florian Götzinger et al. "COVID-19 in children and adolescents in Europe: a multinational, multicentre cohort study". A: *The Lancet Child & Adolescent Health* 4.9 (2020), pàg. 653 - 661. ISSN: 2352-4642. DOI: https://doi.org/10.1016/S2352-4642(20)30177-2. URL: https://www.sciencedirect.com/science/article/pii/S2352464220301772.

[75] Manfred S Green et al. "When is an epidemic an epidemic?" A: *The Israel Medical Association journal: IMAJ* 4.1 (2002), pàg. 3 - 6.

[76] Volker Grimm. "Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future?" A: *Ecological modelling* 115.2-3 (1999), pàg. 129 - 148.

[77] Wei-jie Guan et al. "Clinical characteristics of coronavirus disease 2019 in China". A: *New England journal of medicine* 382.18 (2020), pàg. 1708 - 1720.

[78] Wei-jie Guan et al. "Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis". A: *European Respiratory Journal* 55.5 (2020). ISSN: 0903-1936. DOI: 10.1183/13993003.00547-2020. eprint: https://erj.ersjournals.com/content/55/5/2000547.full.pdf. URL: https://erj.ersjournals.com/content/55/5/2000547.

[79] Bingqi Guo, X George Xu i Chengyu Shi. "Real time 4D IMRT treatment planning based on a dynamic virtual patient model: proof of concept". A: *Medical physics* 38.5 (2011), pàg. 2639 - 2650.

[80] Giorgio Guzzetta et al. "Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting". A: *Journal of theoretical biology* 289 (2011), pàg. 197 - 205.

[81] James W Haefner. *Modeling Biological Systems:: Principles and Applications*. Springer Science & Business Media, 2005.

[82] Emma B Hodcroft et al. “Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020”. A: *MedRxiv* (2020).

[83] Zixin Hu et al. *Artificial Intelligence Forecasting of Covid-19 in China*. 2020. arXiv: 2002.07112 [q-bio.OT].

[84] Chaolin Huang et al. “6-month consequences of COVID-19 in patients discharged from hospital: a cohort study”. A: *The Lancet* (2021).

[85] Chaolin Huang et al. “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China”. A: *The lancet* 395.10223 (2020), pàg. 497 - 506.

[86] Lihan Huang. “A new mechanistic growth model for simultaneous determination of lag phase duration and exponential growth rate and a new Bělehrádek-type model for evaluating the effect of temperature on growth rate”. A: *Food Microbiology* 28.4 (2011). Predictive Modeling in Foods, pàg. 770 - 776. ISSN: 0740-0020. DOI: https://doi.org/10.1016/j.fm.2010.05.019. URL: https://www.sciencedirect.com/science/article/pii/S0740002010001255.

[87] John P.A. Ioannidis, Sally Cripps i Martin A. Tanner. “Forecasting for COVID-19 has failed”. A: *International Journal of Forecasting* (2020). ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2020.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0169207020301199.

[88] Talha N Jilani et al. “Active Tuberculosis”. A: (2018).

[89] Iolanda Jordan et al. “Transmission of SARS-CoV-2 infection among children in summer schools applying stringent control measures in Barcelona, Spain”. A: *Clinical Infectious Diseases* (2021).

[90] Rachel E Jordan, Peymane Adab i K K Cheng. “Covid-19: risk factors for severe disease and death”. A: *BMJ* 368 (2020). DOI: 10.1136/bmj.m1198. eprint: https://www.bmj.com/content/368/bmj.m1198.full.pdf. URL: https://www.bmj.com/content/368/bmj.m1198.

[91] S Kiazyk i TB Ball. “Tuberculosis (TB): Latent tuberculosis infection: An overview”. A: *Canada Communicable Disease Report* 43.3-4 (2017), pàg. 62.

[92] Hiroko Kitaoka, Ryuji Takaki i Béla Suki. “A three-dimensional model of the human airway tree”. A: *Journal of Applied Physiology* 87.6 (1999), pàg. 2207 - 2217.

[93] Shigeru Kondo i Takashi Miura. “Reaction-diffusion model as a framework for understanding biological pattern formation”. A: *science* 329.5999 (2010), pàg. 1616 - 1620.

[94] Jan-Ulrich Kreft et al. “Mighty small: observing and modeling individual microbes becomes big science”. A: *Proceedings of the National Academy of Sciences* 110.45 (2013), pàg. 18027 - 18028.

[95] Andreas Kronbichler et al. “Asymptomatic patients as a source of COVID-19 infections: A systematic review and meta-analysis”. A: *International Journal of Infectious Diseases* 98 (2020), pàg. 180 - 186. ISSN: 1201-9712. DOI: https://doi.org/10.1016/j.ijid.2020.06.052. URL: https://www.sciencedirect.com/science/article/pii/S1201971220304872.

[96] Stephen A Lauer i Kyra H Grantz. “Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application”. A: *Annals of internal medicine* 172.9 (2020), pàg. 577 - 582.

[97] Ramanan Laxminarayan et al. “Global Investments In TB Control: Economic Benefits: As a successful treatment spreads in countries that are the hardest hit by TB, researchers can determine where limited funds can be spent most effectively.” A: *Health affairs* 28.Suppl1 (2009), w730 - w742.

[98] Matthew Le et al. “Neural Relational Autoregression for High-Resolution COVID-19 Forecasting.” A: *Facebook AI* (2020).

[99] Phenyo E Lekone i Bärbel F Finkenstädt. “Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study”. A: *Biometrics* 62.4 (2006), pàg. 1170 - 1177.

[100] Andrew T Levin et al. “Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications”. A: *European journal of epidemiology* (2020), pàg. 1 - 16.

[101] Xing Li et al. “Is hydroxychloroquine beneficial for COVID-19 patients?” A: *Cell death & disease* 11.7 (2020), pàg. 1 - 6.

[102] Robert Loddenkemper et al. “History of tuberculosis”. A: *Tuberculosis (ERS Monograph). Sheffield, European Respiratory Society* 1 (2018), pàg. 08 - 27.

[103] Denis Y Logunov et al. “Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia”. A: *The Lancet* 397.10275 (2021), pàg. 671 - 681.

[104] Cayetana López. “Data Bases Integration and Model Assessment for Covid-19 Prediction and Visualization with AI”. Bachelor’s Thesis. Universitat Politècnica de Catalunya, 2021.

[105] Elisabeth Mahase. “Covid-19: What do we know about “long covid”?” A: *bmj* 370 (2020).

[106] Swapna Mandal et al. “‘Long-COVID’: a cross-sectional study of persisting symptoms, biomarker and imaging abnormalities following hospitalisation for COVID-19”. A: *Thorax* 76.4 (2021), pàg. 396 - 398.

[107] Simeone Marino, Mohammed El-Kebir i Denise Kirschner. “A hybrid multi-compartment model of granuloma formation and T cell priming in tuberculosis”. A: *Journal of theoretical biology* 280.1 (2011), pàg. 50 - 62.

[108] Simeone Marino i Denise E Kirschner. “A multi-compartment hybrid computational model predicts key roles for dendritic cells in tuberculosis infection”. A: *Computation* 4.4 (2016), pàg. 39.

[109] Elisa Martró Català et al. *Analysis and prediction of COVID-19 for EU-EFTA-UK and other countries, report 233*. Inf. tèc. Universitat Politècnica de Catalunya, 2020. URL: `http://hdl.handle.net/2117/346361`.

[110] Edouard Mathieu et al. “A global database of COVID-19 vaccinations”. A: *Nature Human Behaviour* (2021), pàg. 1 - 7.

[111] Perc Matjaž et al. “Forecasting covid-19”. A: *Front. Phys* 8 (2020), pàg. 127.

[112] Conor McAloon et al. “Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research”. A: *BMJ open* 10.8 (2020), e039652.

[113] David N McMurray. “Guinea pig model of tuberculosis”. A: *Tuberculosis: pathogenesis, protection, and control* (1994), pàg. 135 - 147.

[114] Mandeep R Mehra, Frank Ruschitzka i Amit N Patel. "Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis". A: *The Lancet* 395.10240 (2020), pàg. 1820.

[115] Seyed M Moghadas et al. "Evaluation of COVID-19 vaccination strategies with a delayed second dose". A: *PLoS Biology* 19.4 (2021), e3001211.

[116] Saurabh Mukhopadhyay i Linda M. Haines. "Bayesian D-optimal designs for the exponential growth model". A: *Journal of Statistical Planning and Inference* 44.3 (1995), pàg. 385 - 397. ISSN: 0378-3758. DOI: https://doi.org/10.1016/0378-3758(94)00056-2. URL: https://www.sciencedirect.com/science/article/pii/0378375894000562.

[117] César Muñoz-Fontela et al. "Animal models for COVID-19". A: *Nature* 586.7830 (2020), pàg. 509 - 515.

[118] Patrick R. Murray, Ken S. Rosenthal i Michael A. Pfaller. *Medical Microbiology*. Elsevier, 2016.

[119] National Institute for Public Health and the Environment Ministry of Health, Welfare and Sport. *Variants of the coronavirus SARS-CoV-2*. https://www.rivm.nl/en/coronavirus-covid-19/virus-sars-cov-2/variants. Online; accessed 1 June 2021. 2021.

[120] National Library of Medicine. *PubMed.gov COVID-19 2020*. https://pubmed.ncbi.nlm.nih.gov/?term=covid-19&filter=years.2020-2020. Online; accessed 10 May 2021. 2021.

[121] National Library of Medicine. *PubMed.gov Tuberculosis 2020*. https://pubmed.ncbi.nlm.nih.gov/?term=tuberculosis&filter=years.2020-2020. Online; accessed 10 May 2021. 2021.

[122] Tuen Wai Ng, Gabriel Turinici i Antoine Danchin. "A double epidemic model for the SARS propagation". A: *BMC Infectious Diseases* 3.1 (2003), pàg. 1 - 16.

[123] John N Nkengasong et al. *COVID-19 vaccines: how to ensure Africa has access*. 2020.

[124] Maria Helena Santos de Oliveira, Giuseppe Lippi i Brandon Michael Henry. "Sudden rise in COVID-19 case fatality among young and middle-aged adults in the south of Brazil after identification of the novel B. 1.1. 28.1 (P. 1) SARS-CoV-2 strain: analysis of data from the state of Parana". A: *medRxiv* (2021).

[125] World Health Organization. *Global tuberculosis report 2020*. Inf. tèc. World Health Organization, 2020.

[126] World Health Organization. *Novel Coronavirus – China*. Inf. tèc. World Health Organization, gen. de 2020.

[127] World Health Organization et al. *Pandemic influenza preparedness and response: a WHO guidance document*. World Health Organization, 2009.

[128] World Health Organization. *Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003*. Inf. tèc. World Health Organization, des. de 2003.

[129] World Health Organization. *TB: a global emergency*. Inf. tèc. World Health Organization, 1994.

[130] Moises Palaci et al. "Cavitary disease and quantitative sputum bacillary load in cases of pulmonary tuberculosis". A: *Journal of clinical microbiology* 45.12 (2007), pàg. 4064 - 4066.

[131] Yu Pang et al. "Epidemiology of extrapulmonary tuberculosis among inpatients, China, 2008–2017". A: *Emerging infectious diseases* 25.3 (2019), pàg. 457.

[132] Tom Paulson. "Epidemiology: a mortal foe". A: *Nature* 502.7470 (2013), S2 - S3.

[133] Hannah Peckham et al. "Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission". A: *Nature communications* 11.1 (2020), pàg. 1 - 10.

[134] Mathias Peirlinck et al. "Visualizing the invisible: The effect of asymptomatic transmission on the outbreak dynamics of COVID-19". A: *Computer Methods in Applied Mechanics and Engineering* 372 (2020), pàg. 113410.

[135] Bruce Pell et al. "Using phenomenological models for forecasting the 2015 Ebola challenge". A: *Epidemics* 22 (2018), pàg. 62 - 70.

[136] Delphine Planas et al. "Sensitivity of infectious SARS-CoV-2 B. 1.1. 7 and B. 1.351 variants to neutralizing antibodies". A: *Nature medicine* (2021), pàg. 1 - 8.

[137] Fernando P Polack et al. "Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine". A: *New England Journal of Medicine* 383.27 (2020), pàg. 2603 - 2615.

[138] Marina Pollán et al. “Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study”. A: *The Lancet* 396.10250 (2020), pàg. 535 - 544.

[139] Marti Pons-Òdena et al. “COVID-19 and respiratory support devices”. A: *Paediatric Respiratory Reviews* 35 (2020), pàg. 61 - 63. ISSN: 1526-0542. DOI: https://doi.org/10.1016/j.prrv.2020.06.015. URL: https://www.sciencedirect.com/science/article/pii/S1526054220301007.

[140] Clara Prats et al. “Local Inflammation, Dissemination and Coalescence of Lesions Are Key for the Progression toward Active Tuberculosis: The Bubble Model”. A: *Frontiers in Microbiology* 7 (2016), pàg. 33. ISSN: 1664-302X. DOI: 10.3389/fmicb.2016.00033. URL: https://www.frontiersin.org/article/10.3389/fmicb.2016.00033.

[141] Clara Prats et al. “Local inflammation, dissemination and coalescence of lesions are key for the progression toward active tuberculosis: The bubble model”. A: *Frontiers in microbiology* 7 (2016), pàg. 33.

[142] United Nations Development Programme. *Human Development Report 2020*. Inf. tèc. United Nations Development Programme, 2020.

[143] Public Health England. *GOV.UK Coronavirus (COVID-19) in the UK*. https://coronavirus.data.gov.uk/details/download. Online; accessed 13 May 2021. 2021.

[144] Manaf Al-Qahtani et al. “The prevalence of asymptomatic and symptomatic COVID-19 in a cohort of quarantined subjects”. A: *International Journal of Infectious Diseases* 102 (2021), pàg. 285 - 288. ISSN: 1201-9712. DOI: https://doi.org/10.1016/j.ijid.2020.10.091. URL: https://www.sciencedirect.com/science/article/pii/S1201971220322943.

[145] RAC1. *A Catalunya hi hauria entre 600.000 i un milió de casos reals de coronavirus*. https://www.rac1.cat/programes/via-lliure/20200502/48896685128/clara-prats-ivestigadora-upc-contagis-coronavirus-catalunya-600000-milio-casos-reals.html. Online; published 2 May 2020. 2020.

[146] Romain Ragonnet et al. “Revisiting the natural history of pulmonary tuberculosis: a bayesian estimation of natural recovery and mortality rates”. A: *BioRxiv* (2019), pàg. 729426.

[147] Iman Rahimi, Fang Chen i Amir H Gandomi. "A review on COVID-19 forecasting models". A: *Neural Computing and Applications* (2021), pàg. 1 - 11.

[148] Laylaa Ramos et al. "The minipig as an animal model to study Mycobacterium tuberculosis infection and natural transmission". A: *Tuberculosis* 106 (2017), pàg. 91 - 98.

[149] Shubhadeep Roychoudhury et al. "Viral Pandemics of the Last Four Decades: Pathophysiology, Health Impacts and Perspectives". A: *International journal of environmental research and public health* 17.24 (2020), pàg. 9411.

[150] Jerald Sadoff et al. "Safety and efficacy of single-dose Ad26. COV2. S vaccine against Covid-19". A: *New England Journal of Medicine* (2021).

[151] Masaya M Saito et al. "Extension and verification of the SEIR model on the 2009 influenza A (H1N1) pandemic in Japan". A: *Mathematical biosciences* 246.1 (2013), pàg. 47 - 54.

[152] Kankan Sarkar, Subhas Khajanchi i Juan J Nieto. "Modeling and forecasting the COVID-19 pandemic in India". A: *Chaos, Solitons & Fractals* 139 (2020), pàg. 110049.

[153] AF Schalk. "An apparently new respiratory disease of baby chicks". A: *J. Am. Vet. Med. Assoc.* 78 (1931), pàg. 413 - 423.

[154] Sally Sharpe et al. "Ultra low dose aerosol challenge with Mycobacterium tuberculosis leads to divergent outcomes in rhesus and cynomolgus macaques". A: *Tuberculosis* 96 (2016), pàg. 1 - 12.

[155] Antoni Soriano-Arandes et al. "Household SARS-CoV-2 transmission and children: a network prospective study." A: *Clinical Infectious Diseases: an Official Publication of the Infectious Diseases Society of America* (2021).

[156] Chandrashekhar T Sreeramareddy et al. "Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature". A: *BMC infectious diseases* 9.1 (2009), pàg. 1 - 10.

[157] Statem Serum Institut. *EuroMOMO, European mortality monitoring activity.* `https://www.euromomo.eu/`. Online; accessed 26 May 2021. 2021.

[158] William W Stead. "The origin and erratic global spread of tuberculosis: how the past explains the present and is the key to the future". A: *Clinics in chest medicine* 18.1 (1997), pàg. 65 - 77.

[159] Erin K Stokes et al. “Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020”. A: *Morbidity and Mortality Weekly Report* 69.24 (2020), pàg. 759.

[160] Masja Straetemans et al. “Assessing tuberculosis case fatality ratio: a meta-analysis”. A: *PloS one* 6.6 (2011), e20755.

[161] K Styblo, J Meijer, I Sutherland et al. “The transmission of tubercle bacilli. Its trend in a human population.” A: *Selected Papers of the Royal Netherlands Tuberculosis Association* 13 (1971), pàg. 5 - 104.

[162] J. Sznitman. “Respiratory microflows in the pulmonary acinus.” A: *Journal of biomechanics* 46 2 (2013), pàg. 284 - 98.

[163] Ajit K Thakur. “Model: mechanistic vs empirical”. A: *New trends in pharmacokinetics*. Springer, 1991, pàg. 41 - 51.

[164] *The impact of COVID-19 on the TB epidemic: A community perspective*. Inf. tèc. global civil society i TB affected community, 2020.

[165] Kamala Thiagarajan. “Why is India having a covid-19 surge?” A: *BMJ* 373 (2021). DOI: 10.1136/bmj.n1124. eprint: https://www.bmj.com/content/373/bmj.n1124.pdf. URL: https://www.bmj.com/content/373/bmj.n1124.

[166] Nikolaos M Tsoukias i Steven C George. “A two-compartment model of pulmonary nitric oxide exchange dynamics”. A: *Journal of Applied Physiology* 85.2 (1998), pàg. 653 - 666.

[167] DAJ Tyrrell i ML Bynoe. “Cultivation of a novel type of common-cold virus in organ cultures”. A: *British medical journal* 1.5448 (1965), pàg. 1467.

[168] Universitat Politècnica de Catalunya. *Espanya tindria actualment 124.300 casos actius de COVID-19, segons el model matemàtic desenvolupat per la UPC i l'IGTP*. https://www.upc.edu/ca/sala-de-premsa/noticies/espanya-tindria-actualment-124300-casos-actius-de-covid-19-segons-el-model-matematic-desenvolupat-per-la-upc-i-igtp. Online; published 22 May 2020. 2020.

[169] Tomas Urdiales. “The association between ethnic background, socio-economic deprivation and COVID-19 testing, status and health outcomes: a multi-state cohort analysis”. Bachelor's Thesis. Universitat Politècnica de Catalunya, 2021.

[170] P Van Dyck et al. "Imaging of pulmonary tuberculosis". A: *European radiology* 13.8 (2003), pàg. 1771 - 1785.

[171] Eleftheria Vasileiou et al. "Effectiveness of first dose of COVID-19 vaccines against hospital admissions in Scotland: national prospective cohort study of 5.4 million people". A: *Available at SSRN 3789264 (pre-print)* (2021). DOI: http://dx.doi.org/10.2139/ssrn.3789264.

[172] Marina Vegué. "Model tridimensional de l'arbre bronquial huma per a l'estudi de la disseminació de mycobacterium tuberculosis". Treb. fin. de màst. Universitat de Barcelona, 2012.

[173] Ravi A Vijayendran, Frances S Ligler i Deborah E Leckband. "A computational reaction- diffusion model for the analysis of transport-limited kinetics". A: *Analytical Chemistry* 71.23 (1999), pàg. 5405 - 5412.

[174] D Visca et al. "Tuberculosis and COVID-19 interaction: a review of biological, clinical and public health effects". A: *Pulmonology* (2021).

[175] Merryn Voysey et al. "Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK". A: *The Lancet* 397.10269 (2021), pàg. 99 - 111. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(20)32661-1. URL: https://www.sciencedirect.com/science/article/pii/S0140673620326611.

[176] Bolin Wang et al. "Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis". A: *Aging (Albany NY)* 12.7 (2020), pàg. 6049.

[177] Pengfei Wang et al. "Increased resistance of SARS-CoV-2 variant P. 1 to antibody neutralization". A: *Cell host & microbe* (2021).

[178] Ewald R Weibel, André Frédérick Cournand i Dickinson W Richards. *Morphometry of the human lung*. Vol. 1. Springer, 1963.

[179] Howard Howie Weiss. "The SIR model and the foundations of public health". A: *Materials matematics* (2013), pàg. 0001 - 17.

[180] Global Tuberculosis Programme WHO TEAM. *Impact of the COVID-19 pandemic on TB detection and mortality in 2020*. Inf. tèc. World Health Organization, 2021.

[181] World Health Organization. *Timeline: WHO's COVID-19 response*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline. Online; accessed 10 May 2021. 2021.

[182] World Health Organization. *WHO Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/WHO-COVID-19-global-data.csv. Online; accessed 13 May 2021. 2021.

[183] World Health Organization. *WHO TB burden estimates*. https://extranet.who.int/tme/generateCSV.asp?ds=estimates. Online; accessed 13 May 2021. 2021.

[184] Yi-Chi Wu, Ching-Sung Chen i Yu-Jiun Chan. "The outbreak of COVID-19: an overview". A: *Journal of the Chinese medical association* 83.3 (2020), pàg. 217.

[185] Qing Yang et al. "Just 2% of SARS-CoV-2- positive individuals carry 90% of the virus circulating in communities". A: *Proceedings of the National Academy of Sciences* 118.21 (2021).

[186] Fei Zhou et al. "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study". A: *The lancet* 395.10229 (2020), pàg. 1054 - 1062.

[187] Philip Ziegler. *The black death*. Faber & Faber, 2013.

Part IV.
Appendix

A. Glossary

B. List of publications and research activity

A Glossary

Acronyms

AbM	Agent-based Model
ATB	Active Tuberculosis
AQuAS	*Agència de Qualitat i Avaluació Sanitàries de Catalunya*
BCG	Bacillus Calmette-Guérin
BIOCOM-SC	Computational Biology and Complex Systems
CD	Communicable Disease
CFR	Case Fatality Rate
CMCiB	Comparative Medicine and Bioimage Centre of Catalonia
COVID-19	Coronavirus Disease 2019
EMA	European Medicines Agency
EPG	Effective Potential Growth
FDA	Food and Drug Administration
GDP	Gross Domestic Product
HDI	Human Development Index
IbM	Individual-based Model
ICS	*Insitut Català de la Salut*
IFR	Infection Fatality Rate
IGTP	Germans Trias i Pujol Research Institute
LTBI	Latent Tuberculosis Infection
MDR-TB	Multidrug-resistant Tuberculosis
Mtb	*Mycobacterium tuberculosis*
NCD	Non-communicable Disease
NPI	Non-pharmaceutical Interventions
SARS-Cov-2	Severe Acute Respiratory Syndrome Coronavirus 2

SN	Sputum smear negative
SP	Sputum smear positive
TB	Tuberculosis
UPC	*Universitat Politècnica de Catalunya*
UTE	*Unitat de Tuberculosi Experimental*
WHO	World Health Organization

B List of publications and research activity

List of publications

Included in this thesis:

- **Català M**, Prats C, López D, Cardona PJ, Alonso S. A reaction-diffusion model to understand granulomas formation inside secondary lobule during tuberculosis infection. *PLOS One*. 2020 Sep 16;15(9):e0239289.

- **Català M**, Bechini J, Tenesa M, Pérez R, Moya M, Vilaplana C, Valls J, Alonso S, López D, Cardona PJ, Prats C. Modelling the dynamics of tuberculosis lesions in a virtual lung: Role of the bronchial tree in endogenous reinfection. *PLOS Computational Biology*. 2020 May 20;16(5):e1007772.

- Cardona PJ, **Català M**, Prats C. Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance. *Scientific reports*. 2020 Jan 8;10(1):1-20.

- **Català M**, Pino D, Marchena M, Palacios P, Urdiales T, Cardona PJ, Alonso S, López-Codina D, Prats C, Alvarez-Lacalle E. Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers. *PLOS One*. 2021 Jan 7;16(1):e0243701.

- **Català M**, Alonso S, Alvarez-Lacalle E, López D, Cardona PJ, Prats C. Empirical model for short-time prediction of COVID-19 spreading. *PLOS Computational Biology*. 2020 Dec 9;16(12):e1008431.

Other published articles:

- Cardona PJ, **Català M**, Arch M, Arias L, Alonso S, Cardona P, López D, Vilaplana C, Prats C. Can systems immunology lead tuberculosis eradication?. *Current opinion in systems biology*. 2018 Dec 1;12:53-60.

- Arias L, Cardona P, **Català M**, Campo-Pérez V, Prats C, Vilaplana C, Julián E, Cardona PJ. Cording Mycobacterium tuberculosis Bacilli Have a Key Role in the Progression towards Active Tuberculosis, Which is Stopped by Previous Immune Response. *Microorganisms*. 2020 Feb;8(2):228.

- White AD, Sarfas C, Sibley LS, Gullick J, Clark S, Rayner E, Gleeson F, **Català M**, Nogueira I, Cardona PJ, Vilaplana C. Protective Efficacy of Inhaled BCG Vaccination Against Ultra-Low Dose Aerosol M. tuberculosis Challenge in Rhesus Macaques. *Pharmaceutics*. 2020 May;12(5):394.

- Sepúlveda M, Llufriu S, Martínez-Hernández E, **Català M**, Artola M, Hernando A, Montejo C, Pulido-Valdeolivas I, Martínez-Heras E, Guasp M, Solana E. Incidence and Impact of COVID-19 in MS: A Survey From a Barcelona MS Unit. *Neurology-Neuroimmunology Neuroinflammation*. 2021 Mar 4;8(2).

- Pérez-Arnal R, Conesa D, Alvarez-Napagao S, Suzumura T, **Català M**, Alvarez-Lacalle E, Garcia-Gasulla D. Comparative Analysis of Geolocation Information through Mobile-Devices under Different COVID-19 Mobility Restriction Patterns in Spain. *ISPRS International Journal of Geo-Information*. 2021 Feb;10(2):73.

- Jordan I, de Sevilla MF, Fumado V, Bassat Q, Bonet-Carne E, Fortuny C, Garcia-Miquel A, Jou C, Adroher C, Casas MM, Girona-Alarcon M, **Català M** et al. Transmission of SARS-CoV-2 infection among children in summer schools applying stringent control measures in Barcelona, Spain. *Clinical Infectious Diseases*. 2021 Mar 12.

- **Català M**, Marchena M, Conesa D, Palacios P, Urdiales T, Alonso S, Alvarez E, Lopez D, Cardona P-J, Prats C. Monitoring and analysis of COVID-19 pandemic: the need for an empirical approach. *Frontiers in Public Health*. 2021 Jul 8;9:806.

- **Català M**, Coma E, Alonso S, Alvarez E, Cordomi S, Lopez D, Fina F, Medina-Peralta M, Prats C,Prieto-Alhambra D. Risk diagrams based on primary care electronic medical records and linked real-time PCR data to monitor local COVID-19 outbreaks during the summer 2020: a prospective cohort study including 7,671,862 people in Catalonia. *Frontiers Public Health*. 2021 Jul 5;9:890.

- López F, Estrada O, Mitjà O, Hernández G, Prat N, Bonet JM, Isnard M, Moreno N, Blanco I, Vilar M, **Català M**, Aran A, Argimon JM, Clotet B, Ara J. A Cost-Benefit Analysis of the COVID-19 Asymptomatic Mass Testing Strategy in the North Metropolitan Area of Barcelona. In submission: *International Journal of Environmental Research and Public Health*. 2021 Jun 30;18(13):7028.

- Alonso S, Alvarez-Lacalle E, **Català M**, Lopez-Codina D, Jordan I, García-García JJ, Soriano-Arandes A, Lazcano U, Gatell A, Cappevila R, Soler-Palacin P, Bassat Q, Prats C. Age-dependency of the propagation rate of COVID-19 inside bubble scholar groups in Catalonia, Spain. *The Pediatric Infectious Disease Journal*. 2021 Jul 27.

- Perramon A, Soriano-Arandes A, Pino D, Lazcano U, Andrés C, **Català M**, Gatell A, Carulla M, Canadell D, Ricós G, Riera-Bosch MT, Burgaya S, Salvadó O, Cantero J, Vilà M, Poblet M, Sánchez A, Ristol AM, Serrano P, Antón A, Prats C, Soler-Palacin P. Schools as a framework for COVID-19 epidemiological surveillance of children in Catalonia, Spain: a population-based study. *Frontiers in Pediatrics*. Accepted.

Articles under review:

- **Català M**, Li X, Prats C, Prieto-Alhambra D. The Impact of Prioritisation and Dosing Intervals on the Effects of COVID-19 Vaccination in Europe: An Agent-Based Cohort Model. *In submission*. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3793540`

- Cardona PJ, **Català M**, Prats C. The origin and maintenance of tuberculosis is explained by its subclinical course, the Neolithic revolution being the trigger for its devastating deadly drift. *In submission*.

- Melé-Casas M, F. de Sevilla M, Hernández-García M, Pons-Tomàs G, Launes C, Bassat Q, Fumadó V, Fortuny C, García-Miquel A, Bonet-Carne E, **Català M** et al. Low transmission of SARS-CoV-2 derived from children in family clusters: an observational study of family households in the Barcelona Metropolitan Region, Spain. *In submission*.

Articles under preparation (drafts):

- **Català M**, Alonso S, Lopez D, Cardona PJ, Prats C. Local small breathing amplitude increases active Tuberculosis development. *In preparation.*
- Villanueva MI, **Català M**, Lopez C, Alonso S, Lopez D, Prats C, Alvarez E. Short-term Gompertz empirical predictions of COVID-19 evolution in Europe improve accuracy upon country-report pattern corrections. *In preparation.*
- Cuadras P, Rafart G, **Català M**, Pérez R, Rosenthal A, Martinson N, Bechini J, Vilaplana C. Generation and validation of a X-ray Score to assess improvement of Pulmonary TB disease in X-rays (BCN-SA X-ray Score). *In preparation.*
- Nogueira I, **Català M**, Prats C, Sharpe S, Vilaplana C, Cardona PJ. On the progression of TB in the non-human primate macaques. A new tool for evaluating TB vaccines. *In preparation.*
- Antoñanzas J, Perramon A, Lopez C, Aguilera C, Boneta M, **Català M**, Capdevila R, Prats C, Soriano A. Symptoms-based predictive model of covid-19 disease in children. *In preparation.*
- Joseph B, Sarrias M, Gonzalez S, **Català M**, Alvarez E, Alonso S, Lopez D, Prats C, Arnaldos M. A methodology for the use of waste water measurements as a sentinel for COVID-19 in Catalonia. A retrospective study. *In preparation.*

Talks

- **Català M**, López D, Valls J, Vilaplana C, Cardona PJ, Bechini J, Pérez R, Tenesa M, Prats C. A 3D computational model for understanding tuberculosis lesions dynamics in lungs: the bridge from experimental to human tuberculosis. *Correlating Clinical and Experimental Pathology.* Poster presentation. Castelló d'Empúries, June 2016.

- **Català M**, Prats C, Vilaplana C. From Latent Infection to Active Disease: new experimental and mathematical approaches in tuberculosis research. *Meeting MycoClub 2016.* Oral presentation. Madrid, June 2016.

- **Català M**, Alonso S, López D, Prats C. Reaction-diffusion model for the understanding of tuberculosis dynamics at the scale of a secondary pulmonary lobule. *FisEs'18 XXII Congreso de Física Estadística.* Poster presentation. Madrid, October 2018.

- **Català M**, Alonso S, López D, Prats C. Reaction-diffusion model for the understanding of tuberculosis dynamics at the scale of a secondary pulmonary lobule. *Jornades de Recerca del Departament de Física de la Universitat Politècnica de Catalunya.* Poster presentation. Barcelona, February 2019.

- Cardona PJ, **Català M**, Prats C. The extraordinary impact of Mycobacterium tuberculosis in the Paleolithic predicts unprecedented population growth due to female protection against tuberculosis. *European Society of Mycobacteriology Congress.* Poster presentation. Valencia, 2019.

- Cardona PJ, **Català M**, Prats C. The extraordinary impact of Mycobacterium tuberculosis in the Paleolithic predicts unprecedented population growth due to female protection against tuberculosis. *Jornades de Recerca del Departament de Física de la Universitat Politècnica de Catalunya.* Poster presentation. Barcelona, February 2020.

- Alonso S, **Català M**, Prats C, Lopez D, Alvarez E, Marchena M, Cardona PJ. Modelo empírico para la predicción a corto plazo de la propagación de COVID-19. *Inteligencia Artificial y Big Data contra el Covid-19.* Oral presentation. Online, April 2020.

- **Català M**, Cardona PJ, Alonso S, Lopez D, Prats C. Levy flight and local diffusion determine transition to active version of tuberculosis infection. *FisEs'20 XXIII Congreso de Física Estadística*. Poster presentation. Zaragoza, May 2020. Cancelled due to COVID-19 pandemic.

- Prats C, **Català M**, Alonso S, Lopez D, Alvarez E, Marchena M, Cardona PJ. Ús de models per a l'anàlisi i predicció de la dinàmica epidemiològica de la COVID-19. *X Jornada de Enfermedades Emergentes 2020*. Oral presentation. Barcelona, June 2020.

- Alonso S, **Català M**, Prats C, Lopez D, Alvarez E, Marchena M, Cardona PJ. The use of empiric models for the short-term prediction of COVID-19. Risk assessments in epidemiological dynamics. *Jornades SCM, Matemàtiques i Covid-19*. Oral presentation. Online, July 2020.

- **Català M**, Prats C, Cardona PJ, Capdevila S. Comparative Medicine and Bioimage Centre of Catalonia: The new 3Rs reference centre in Spain. *11th World Congress on Alternatives and Animal use in Life sciences*. Poster presentation. Maastricht, August 2020. Cancelled due to COVID-19 pandemic.

- **Català M**, Alonso S, Cardona PJ, López D, Prats C. In silico modelling of Tuberculosis in a computational lung of minipig. *11th World Congress on Alternatives and Animal use in Life sciences*. Poster presentation. Maastricht, August 2020. Cancelled due to COVID-19 pandemic.

- Prats C, **Català M**, Alonso S, Lopez D, Alvarez E. Perspectivas futuras de la COVID-19. *XI Jornada de Enfermedades Emergentes 2021*. Oral presentation. Barcelona, June 2021.

Seminars

- **Català M**. The bubble model, a computational model for undestanding tuberculosis lesions dynamics in lungs. *Seminar in Mathematical and Physical Biology*. Barcelona, October 2017.
- **Català M**. The bubble model, a computational model for undestanding tuberculosis lesions dynamics in lungs. *Physics Engigneering Degree Biophysics I conferences*. Barcelona, May 2017.
- **Català M**, Prats C. Computational modelling towards 3Rs. Case study: tuberculosis in a virtual lung. *3R CMCiB Seminars*. Badalona, December 2019.
- **Català M**, Prats C. Computational modelling in COVID-19 research. *IGTP COVID-19 Seminar*. Online, July 2020.
- **Català M**, Prats C. Grup de Biologia Computacional i Sistemes Complexos (BIOCOM-SC), Ciència en temps de pandèmia. *Fira del Coneixement*. Online, April 2021.

  Coauthor of several seminars in the framework of the diffusion of Computational Biology and Complex Systems research group (BIOCOM-SC) work during the COVID-19 pandemic.

Supervised students

Bachelor thesis (30 ECTS*):

- Estévez A. Machine Learning en proves neurològiques de model de ratolins per l'ictus. Bachelor thesis, Physics Engineering, Universitat Politècnica de Catalunya. 2020. Supervisors: Martí Català, Clara Prats, Octavi Martí, Teresa Gasull.
- Villanueva MI. Identificació i caracterització automatitzada de lesions de tuberculosi en imatges de tomografia axial computada. Bachelor thesis, Physics Engineering, Universitat Politècnica de Catalunya. 2020. Supervisors: Martí Català, Raul Benitez, Clara Prats.
- Urdiales T. The association between ethnic background and COVID-19 testing, status and health outcomes: a multi-state cohort analysis including 406406 participants. Bachelor thesis, Physics Engineering, Universitat Politècnica de Catalunya. 2021. Supervisors: Martí Català, Daniel Prieto-Alhambra, Albert Prats-Uribe, Clara Prats.

Other projects (3 ECTS*):

- Navarro P, Ferrer JM, Solerdelcoll N, Valenzuela A. From Abm to continuous. A Tuberculosis alveolar infection model. PEF2 project, Physics Engineering, Universitat Politècnica de Catalunya. 2018. Supervisor: Martí Català.
- Perramon A, Villanueva MI. Building of a computational bronchial tree. PEF2 project, Physics Engineering, Universitat Politècnica de Catalunya. 2019. Supervisor: Martí Català.
- Pradas S, Rovira D, Simó X. From COVID-19 analysis using the Gompertz function. PEF2 project, Physics Engineering, Universitat Politècnica de Catalunya. 2020. Supervisors: Martí Català, Clara Prats.
- Dot L, Kirchner H. Health care system dynamics modelization for COVID-19. PEF2 project, Physics Engineering, Universitat Politècnica de Catalunya. 2020. Supervisors: Martí Català, Enric Alvarez-Lacalle, Daniel López.
- Urdiales T, Palacios P. Global analysis of Covid-19 in Europe. PEF2 project, Physics Engineering, Universitat Politècnica de Catalunya. 2020. Supervisors: Martí Català, Clara Prats.

* One ECTS (European Credit Transfer System) is equivalent to 25 working hours for the student.

Collaborations

- Pere-Joan Cardona, Cristina Vilaplana, Paula Cardona, Albert Despuig, Lilibeth Arias. Tuberculosis Experimental Unit, Institute for Health Science Research Germans Trias i Pujol. Several common projects about TB, our main support on the clinical expertise in TB. 2015-present.
- Isabel Nogueira, Jordi Bechini, Patricia Cuadras, Ricard Pérez, Montserrat Tenesa. Radiodiagnostic department, Hospital Universitari Germans Trias i Pujol. They identified lesions of minipigs, macaques and goats Tuberculosis lesions in Computational Tomography scans. 2015-present.
- Teresa Gasull, Octavi Martí. Cellular & Molecular Neurobiology, Institute for Health Science Research Germans Trias i Pujol. Using Artificial Intelligence techniques to determine mice ictus lesions severity. 2019-present.
- Cristina Colls, Borja Velasco, Didier Dominguez. Agència de Qualitat i Avaluació Sanitàries de Catalunya. Methodology development of COVID-19 indicators and predictions during COVID-19 pandemic. 2020-present.
- Manolo Medina, Ermengol Coma, Silvia Cordomí. Institut Català de la Salut. Methodology development of COVID-19 indicators and predictions during COVID-19 pandemic and the study of COVID-19 clinic diagnoses. 2020-present.
- Daniel Prieto-Alhambra, Albert Prats, Xintong Li. Nuffield Department of Orthopaedics, Rheumatology and Musculoeskeletal Sciences. Studying COVID-19 clinic diagnoses, vaccination strategies and ethnic differences. 2020-present
- Iolanda Jordan, Juan José García, Quique Bassat, Eduard Gratacós. Hospital Sant Joan de Déu, Kids Corona project. Study of kids COVID-19 transmission in summer schools. 2020.
- Antoni Soriano (Vall d'Hebron hospital) and COPEDI-CAT pediatricians network. Study the transmission of COVID-19 in children and their symptoms. 2020-present.
- Maximilian Nickel, Alex Pompe. Facebook Artificial Intelligence. Developing Artificial Intelligence methods to predict COVID-19 evolution in different European regions. 2020-present.

- Marina Arnaldos, Bernat Joseph, Miquel Sàrrias, Silvia Gonzalez. Cetaqua, Water Technology Centre. Study the correlation between COVID-19 presence in residual water and this disease incidence in the population. 2020-present.
- Raquel Pérez, Sergio Alvarez, Dario Garcia. Barcelona Supercomputing Centre. Study the correlation between Google and Facebook mobility indexes with the evolution of COVID-19 epidemic. 2020.
- Albert Saiz. Multiple Sclerosis unit, Hospital Clínic Barcelona. Study the COVID-19 incidence in the Multiple Sclerosis units patients. 2020.
- Francesc Lopez. Health Economy, Hospital Universitari Germans Trias i Pujol. Economic study on the mass testing strategy in Catalonia. 2021.
- Roger Paredes, Marc Noguera, Ignacio Blanco, Pere-Joan Cardona. COVID-19 sequencing, IRSICaixa and HUGTiP. Follow COVID-19 variants percentage evolutions. 2021-present.
- Allisson Dantas, Edneide Ramalho. Laboratório de Imunopatologia Keizo Asami, Universidade Federal de Pernambuco. Study of malaria in Brazil and COVID-19 collaboration on the diffusion of BIOCOM-SC methodology. 2018-present.
- Neil Martinson. Perinatal HIV Research Unit, University of the Witwatersrand, Soweto, South Africa. Development of a score to monitor the improvement of chest X-ray image associated with active TB (BCN-SA score). 2018-present.
- Sally Sharpe. Public Health England. They provided us the results of macaques active tuberculosis experiments to analyse them. 2019-2020.
- Nura Ahmad. Gombe State University, Nigeria. Study Tuberculosis spreading with an Agent Based Model. 2018-present.