Title: Risk of COVID-19 hospitalisation rises exponentially with age, inversely proportional to T-cell production

Authors: Sam Palmer[1]*, Nik Cunniffe[2], Ruairí Donnelly[2].

Affiliations:

[1]Mathematical Institute, University of Oxford, Oxford, UK.

[2]Department of Plant Sciences, University of Cambridge, Cambridge, UK

*Correspondence to: palmer@maths.ox.ac.uk

Abstract

Here we report that COVID-19 hospitalisation rates follow an exponential relationship with age, doubling for every 16 years of age or equivalently increasing by 4.5% per year of life (R^2=0.98). This mirrors the well studied exponential decline of both thymus volume and T-cell production, which halve every 16 years. COVID-19 can therefore be added to the list of other diseases with this property, including those caused by MRSA, West Nile virus, Streptococcus Pneumonia and certain cancers, such as chronic myeloid leukemia and brain cancers. In addition, incidence of severe disease and mortality due to COVID-19 are both higher in men, consistent with the degree to which thymic involution (and the decrease in T-cell production with age) is more severe in men compared to women. For under 20s, COVID-19 incidence is remarkably low. A Bayesian analysis of daily hospitalisations, accounting for contact-based and environmental transmission, indicates that non-adults are the only age group to deviate significantly from the exponential relationship. Our model fitting suggests under 20s have 49-75% additional immune protection beyond that predicted by strong thymus function alone, consistent with increased juvenile cross-immunity from other viruses. We found no evidence for differences between age groups in susceptibility to overall infection, or, relative infectiousness to others. The strikingly simple inverse relationship between COVID-19 risk and thymic T-cell output reported here begs a mechanistic understanding and suggests that T-cell based therapies may be a promising target.

Introduction

Epidemiological patterns in the incidence of a disease can provide insight into the mechanisms of disease progression[1–4]. The degradation of the adaptive immune system with age is already acknowledged to be a major risk factor for both infectious and non-infectious diseases and may play a role in understanding the emerging COVID-19 epidemic. Thymus volume, and the concomitant production of T-cells, decrease exponentially with age with a half-life of 16 years, or equivalently by 4.5% per year[5,6]. These changes in the adaptive immune system lead to less robust immune responses in elderly individuals[7]. In this paper, we analyse age and gender trends in national COVID-19 hospitalisation data, in order to investigate the role of immune function in the ongoing coronavirus pandemic.

COVID-19 disease progression can be characterised by three consecutive phases of increasing severity[8,9]. First, there are mild symptoms such as a dry cough, sore throat and

fever. After this point the majority of cases will undergo spontaneous regression[10]. Second, some patients can develop viral pneumonia, requiring hospitalization[8]. The third stage, typically occurring three weeks after the onset of symptoms, is characterised by fibrosis[8]. This leads to life threatening symptoms including organ failure, septic shock, acute respiratory distress syndrome, encephalitis, cerebrovascular events and decimation of the lymph nodes[10–12]. COVID-19 patients often exhibit lymphopenia, i.e. extremely low blood T-cell levels, even in the first few days after the onset of symptoms, which is a predictor of disease progression and mortality[13,14]. Clinical trials are currently underway to test T-cell based immunotherapies[15] and vaccines that elicit T-cell, as well as antibody, responses[16]. There is evidence that T-cells may be more effective than antibodies as exposed, asymptomatic individuals develop a robust T-cell response without (or before) a measurable humoral response[17].

The relationship between COVID-19 risk and age has been extensively explored[18–20] and age-stratified, contact-based, transmission models have accurately explained various aspects of the pandemic[19–22]. In particular, these studies have found that the risk of severe disease rises with age and is especially low for those under 20. Some studies suggest that non-adults are as likely to be infected as adults, but then have lower risk of disease progression[22] while others find lower risk of both infection and disease progression in the under 20s[20,21]. While these studies have looked at COVID-19 risk and age, here we go further by relating these trends to thymic involution and T-cell production. This may lead to a mechanistic understanding of disease progression.

Several diseases have risk profiles that increase exponentially with age, doubling every 16 years, i.e. risk is proportional to $e^{0.044t}$, where t is age, or equivalently increasing by about 4.5% per year[4]. These diseases are caused by a range of pathogens, from bacterial (MRSA, *S. Pneumonia*) to viral (West Nile virus) and even include some cancers (chronic myeloid leukemia, heart and brain cancers). Since thymus volume and T-cell production both decrease with age exponentially, halving every 16 years[5], disease risk is therefore inversely proportional to T-cell production for these diseases. Consistently, the gender bias in T-cell production also roughly matches the gender bias in disease risk, with men having approx. 1.3-1.5 times higher overall cancer and infectious disease risk[23–25] and approx. 1.5±0.3 times lower T-cell production, as measured by T-cell receptor excision circles (TRECs), a proxy for thymic output[4,6]. As such, fundamental patterns in disease incidence with respect to both age and gender can be directly linked to differences in the adaptive immune system. We therefore tested to see if COVID-19 follows the same trend.

Results

COVID-19 hospitalisation rates

While data on confirmed cases can be highly variable and largely influenced by testing strategies, the data on hospitalisations, which is the focus of this paper, are relatively more reliable. Incidence of COVID-19 hospitalisations, in a number of countries, consistently doubles with every 16 years of age (R^2=0.98 for top three countries, Fig. 1A). Meanwhile, the incidence of all confirmed cases (including mild or asymptomatic) appears roughly constant across adult ages (Fig. S1). One explanation that is consistent with the data is that exposure is approximately uniform for adult age groups and that after exposure, the probability of becoming hospitalised is proportional to $e^{0.044t}$. We will address the age-dependence of exposure in more detail by accounting for assortative social mixing as well as a range of additional age-dependent factors in our Bayesian model (see below).

There is a gender bias in COVID-19 risk, which increases with disease severity (Fig. 1C). This is similar to other diseases, including cancer, where men have 1.33 times the risk of hospitalisation and 1.89 times the risk of death[23,26]. The gender bias in COVID-19 is remarkably similar with a factor 1.35±0.4 for hospitalisation incidence and 1.9±0.4 for mortality (mean ± s.d. Fig. 1C). The slope of the logarithm of the COVID-19 mortality curve is over twice that of the hospitalisation curve, corresponding to an exponential with rate 0.109±0.005 years^{-1} (Fig. 1B). Another way of thinking about the gender bias would be to say that for both hospital incidence and mortality, men are effectively ~6 years older than women in terms of risk. Other risk factors such as BMI can also be viewed similarly to give an individualised effective "Covid age"[27]. The increase in mortality with age may also be explained by comorbidities which increase with age, such as cardiovascular disease, which rises exponentially[28] with a rate of 0.071±0.003 years^{-1}. Since 0.071+0.044=0.115≈0.109, a simple model where the risk of COVID-19 mortality is proportional to risk of cardiovascular disease and inversely proportional to T-cell production would have the correct age-dependence. This would suggest that cardiovascular disease is unlikely to be a risk factor for hospitalisation but could be for subsequent disease progression.

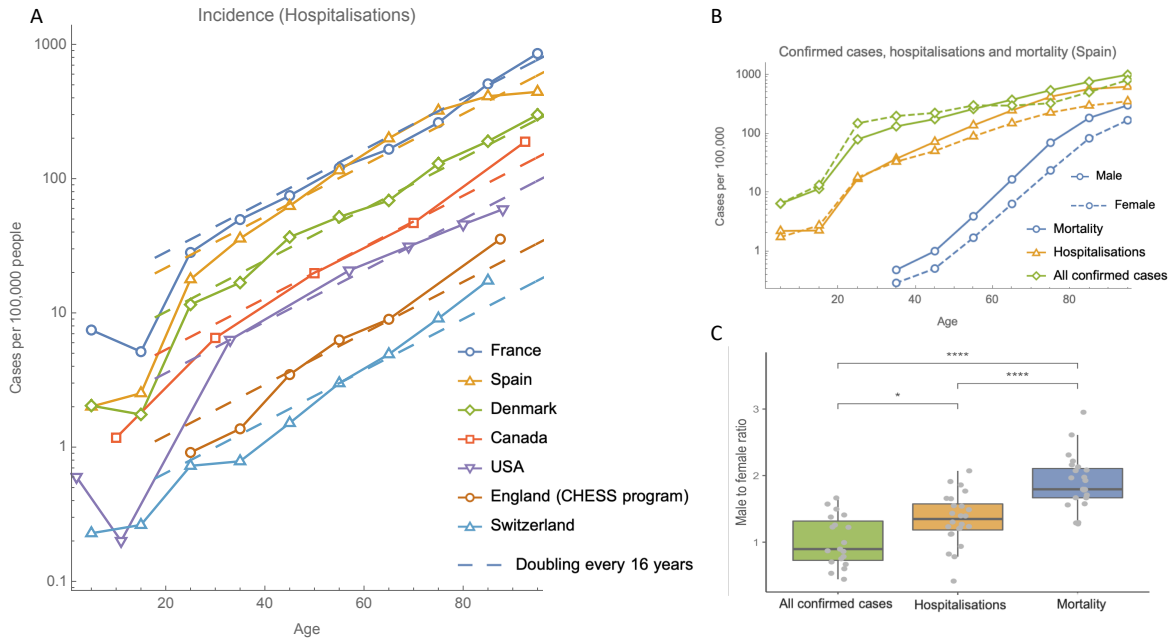


Fig. 1. (A) For adults, incidence of COVID-19 hospitalisations rises exponentially with age, doubling with every 16 years of age. See table S3 for a full list of data sources. (B) Data from Spain on all confirmed cases, hospitalisations and mortality, from a single study early in the epidemic, shows a gender bias which increases with disease severity. (C) Boxplot showing male to female ratios for incidence, hospitalisation rates and mortality, across all age groups with non-zero entries, from the following countries: France, England, Wales and Spain.

Bayesian model

Similar to other diseases, COVID-19 risk is relatively high for very young children (e.g. 0.6 cases per 100,000 for ages 0-4 *vs.* 0.2 cases per 100,000 for ages 5-17 in USA, Fig. 1A). Additionally, older children have a risk lower than expected based on the exponential

increase with age we have identified (Fig. 1A). This is similar to MRSA and *S. Pneumonia* infection, but not West Nile virus (WNV) infection or cancers with similar exponential behaviour [4]. Potential factors underlying the apparent low risk in juveniles include age-dependence in: 1) exposure (e.g. due to heterogeneous social mixing among age groups), 2) disease progression, 3) infection given exposure, and/or, 4) infectiousness to others. Throughout this paper we use the term 'severe infection' synonymously with hospitalisation and we categorise all infections as either mild or severe. In a preliminary analysis, we first incorporated contact matrices into a simple analytically-tractable Susceptible-Infected-Removed (SIR) model to predict the steady state of the age distribution of hospitalisations in France, with the assumption that the probability of severe disease given infection is proportional to $e^{0.044t}$ (see supplementary materials, Fig. S3). This model suggested that age differences in social mixing could, in part, account for the relatively low hospitalisation of non-adults (Fig. S3). However, the other possible factors in low juvenile COVID-19 hospitalisation were not considered in this preliminary analysis.

To incorporate all relevant factors, and to more rigorously test our main hypothesis, we conducted a more detailed analysis of age-dependence based on daily hospitalisation, recovery and death data. We focused on the single country France, for which an unusually comprehensive age distributed dataset is available[19]. All cases in the dataset are either biologically confirmed or present with a computed tomographic image highly suggestive of SARS-CoV-2 infection, and the dataset includes corrections for reporting delays[19]. We formulated an age-structured Bayesian SIR model of infection, partitioning the force of infection into that arising from contacts with mild and severely infected individuals, weighted by age-dependent contact matrices, as well as contact-independent (environmental) transmission. The model fitting exercise focused on inferring a posterior parameter distribution for the probability of severe disease given infection for each age cohort. In addition, posterior distributions were inferred for a range of secondary parameters (Table S2, parameters of the Bayesian analysis), including age-dependent transmissibility and susceptibility.

Our results reiterate that the probability of severe disease given infection increases exponentially with age, at a rate that is remarkably well matched by the rate of thymus decline for all age groups above 20 years (Fig. S4, all adult age groups have 95% credible intervals including the rate of thymus decline). In order to investigate the nature of juvenile deviation from this exponential relationship, we reformulated the analysis to allow deviations from an exponential increase (for the probability of severe disease given infection) for each age cohort (Fig. 2). The posterior parameter distribution for the exponential rate was found to match the rate of thymic degradation (95% CI:0.043-0.053 $years^{-1}$, Fig. 2B). Only the juvenile age-cohort was found to significantly deviate from the exponential response (Fig. 2C), showing a level of additional protection to severe COVID-19 of between 49-75% (Table S1). Our sensitivity analysis allowed – within each age cohort – for deviation from uniform probability of infection given exposure, and, deviation from uniform infectiousness of infected individuals. For both of these we found that none of the age cohorts deviated significantly (in all cases 95% credible intervals included zero deviation, Fig. S5), allowing us to discount these potentially confounding factors.

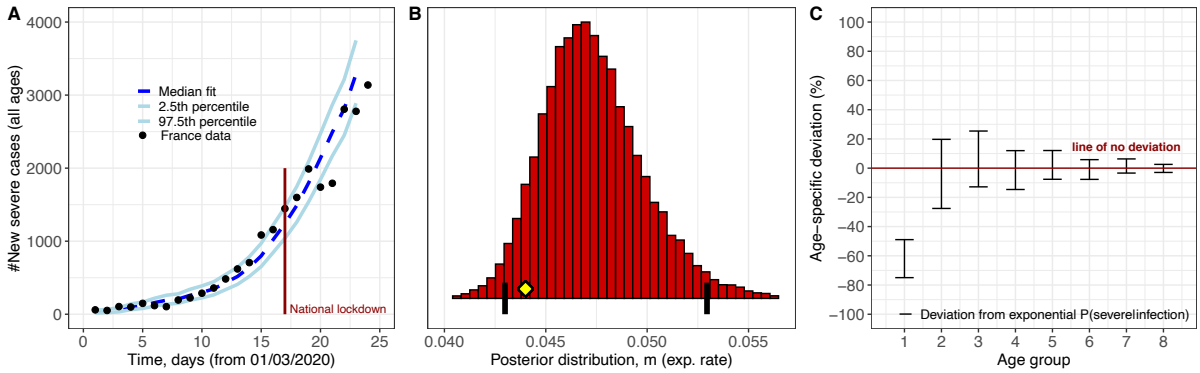


Fig. 2. (A) Forward simulation of the French epidemic using the fitted parameters (B-C) produces a credible interval containing the French hospitalisation data up to day 24 (B) The 95% credible interval for the rate of age dependent exponential growth in hospitalisation probability includes the rate of thymus degradation (0.044 years^{-1}, yellow diamond). Black vertical lines show the 2.5th and 97.5th percentiles. (C) The juvenile cohort has additional significant protection beyond what is predicted by their stronger thymus function (red interval is separated from the zero deviation line for juveniles only). See supplementary information section 'Bayesian modelling' for full description of methods.

The low susceptibility to severe disease given infection in non-adults may be due to cross-protection from other coronaviruses[8,29,30], or even non-specific protection from other respiratory viruses[31], which occur more frequently in non-adults compared to adults[32]. Our estimate of 49-75% protection ties in with a study which found SARS-Cov-2 reactive antibodies in approximately 60% of unexposed individuals aged 6-16 and only 6% in adults[33]. There is also evidence of unexposed individuals having SARS-CoV-2 reactive CD4+ T-cells[34]. Another possible explanation for the low risk in non-adults might come from some intrinsic feature of the immune system. For example, T-cell homeostasis may be maintained differently in the under 20 age group[5]. Intriguingly, the risk of T-lymphoblastic leukemia is approximately constant for adults but highest for ages 5-20 (see ref.[4] Supp. Fig. 5), which may be related to the low disease risk for those ages.

Discussion

Although we have demonstrated a clear relationship between the probability of severe disease and age, it is possible that the relationship is due, in part, to alternative physical processes other than T-cell production. The most closely related cell-type to T-cells are B-cells, which develop in the bone marrow. Bone marrow also shrinks with age, but at a rate that is substantially slower than the thymus[35]. Furthermore, a mechanism for why the probability of hospitalisation is inversely proportional to T-cell production is currently lacking. One possible model features stochastic fluctuations in the number of infected cells and an immune escape threshold which is proportional to T-cell production[4]. This model has the added benefit that it can also explain most of the other (non-exponential) relationships between risk and age seen in various cancer types[4].

Chronic myeloid leukemia (CML) is a type of cancer with an age-dependence remarkably similar to COVID-19. In both diseases the risk of hospitalisation rises exponentially, inversely proportional to T-cell production[4], with gender bias ratios of 1.35±0.4 for COVID-19 and 1.35±0.3 for CML. The mortality risk profiles are also similar (exponential rates: 0.109±0.005 years^{-1} for COVID-19 and 0.103±0.007 years^{-1} for CML, gender bias ratios:1.9±0.4 for COVID-19 and 1.8±0.6 for CML, Fig. S2). CML is

characterised by a single genomic feature, a chromosomal translocation known as the Philadelphia chromosome. This suggests that the probabilities of Philadelphia chromosome formation and COVID-19 infection are approximately age-independent, but that the probabilities of subsequent hospitalisation are T-cell dependent. A good candidate for a potential mechanism involves the phenomenon that increased antigenic load can lead to T-cell exhaustion, characterised by low effector function and clone-specific depletion[36]. T-cell exhaustion is a factor in both cancer and infectious diseases, including COVID-19[37,38], where it has even been shown to be a predictor of mortality[39]. As T-cell production decreases with age, this may lead to an increase in the probability for T-cell exhaustion. In support of this hypothesis, low precursor T-cell numbers have been shown to lead to T-cell exhaustion and disease progression in a mouse cancer model[40]. More specifically, we predict a step in disease progression with a probability exactly inversely proportional to the number of precursor T-cells. When looking at gender biases for COVID-19 hospitalisation and mortality (Fig. 1C) we found factors of 1.35 and 1.9 respectively. We can speculate that since $1.35^2 \approx 1.9$, this might be an indication that among the steps of disease progression, there could be two T-cell dependent steps, one pre-hospitalisation and one post-hospitalisation. The log-slope of the mortality curve being over twice that of the hospitalisation curve is consistent with this hypothesis. One feature of post-hospitalisation disease progression is an IL-6 driven cytokine storm[41], which may be related to T-cell dysfunction[42].

Here we have shown that risk of COVID-19 hospitalisation rises exponentially with age, inversely proportional to T-cell production, in a similar way to several other diseases. Consistently, the gender bias in disease risk also fits this trend. In addition, we found that the under-20 age group benefits from additional protection from severe disease. We hope that these findings will be an important clue in understanding the precise mechanisms involved in disease progression.

References

1. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **91**, 1983–1989 (2004).
2. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
3. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci.* **112**, 118–123 (2015).
4. Palmer, S., Albergante, L., Blackburn, C. C. & Newman, T. J. Thymic involution and rising disease incidence with age. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1883–1888 (2018).
5. Murray, J. M. *et al.* Naive T cells are maintained by thymic output in early ages but by proliferation without phenotypic change after age twenty. *Immunol. Cell Biol.* **81**, 487–495 (2003).
6. Sottini, A. *et al.* Simultaneous Quantification of T-Cell Receptor Excision Circles (TRECs) and K-Deleting Recombination Excision Circles (KRECs) by Real-time PCR. *JoVE J. Vis. Exp.* e52184 (2014) doi:10.3791/52184.
7. Montecino-Rodriguez, E., Berent-Maoz, B. & Dorshkind, K. Causes, consequences, and reversal of immune system aging. *J. Clin. Invest.* **123**, 958–965 (2013).
8. Huang, A. T. *et al.* A systematic review of antibody mediated immunity to coronaviruses: antibody kinetics, correlates of protection, and association of antibody responses with severity of disease. *medRxiv* 2020.04.14.20065771 (2020) doi:10.1101/2020.04.14.20065771.
9. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern. Med.* **180**, 934–943 (2020).

10. Sohrabi, C. *et al.* World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* **76**, 71–76 (2020).
11. Varatharaj, A. *et al.* Neurological and neuropsychiatric complications of COVID-19 in 153 patients: a UK-wide surveillance study. *Lancet Psychiatry* **0**, (2020).
12. Chen, Y. *et al.* The Novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Directly Decimates Human Spleens and Lymph Nodes. *medRxiv* 2020.03.27.20045427 (2020) doi:10.1101/2020.03.27.20045427.
13. Tan, L. *et al.* Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduct. Target. Ther.* **5**, 1–3 (2020).
14. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* (2020) doi:10.1126/science.abc8511.
15. Harrison, C. Coronavirus puts drug repurposing on the fast track. *Nat. Biotechnol.* **38**, 379–381 (2020).
16. Folegatti, P. M. *et al.* Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *The Lancet* **0**, (2020).
17. Sekine, T. *et al.* Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *bioRxiv* 2020.06.29.174888 (2020) doi:10.1101/2020.06.29.174888.
18. T J Newman. *Correlations in US COVID-19 mortality age profiles: epidemic start dates, geography and the PCF hypothesis*. https://zenodo.org/record/3976802 (2020) doi:10.5281/zenodo.3976802.
19. Salje, H. *et al.* Estimating the burden of SARS-CoV-2 in France. *Science* (2020) doi:10.1126/science.abc3517.

308 20. Davies, N. G. *et al.* Age-dependent effects in the transmission and control of COVID-19
309 epidemics. *Nat. Med.* 1–7 (2020) doi:10.1038/s41591-020-0962-9.
310 21. Zhang, J. *et al.* Changes in contact patterns shape the dynamics of the COVID-19
311 outbreak in China. *Science* **368**, 1481–1486 (2020).
312 22. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their
313 close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.*
314 S1473309920302875 (2020) doi:10.1016/S1473-3099(20)30287-5.
315 23. Dorak, M. T. & Karpuzoglu, E. Gender differences in cancer susceptibility: an
316 inadequately addressed issue. *Front. Genet.* **3**, 268 (2012).
317 24. Debacker, M. *et al.* Mycobacterium ulcerans disease: role of age and gender in incidence
318 and morbidity. *Trop. Med. Int. Health* **9**, 1297–1304 (2004).
319 25. Guerra-Silveira, F. & Abad-Franch, F. Sex Bias in Infectious Disease Epidemiology:
320 Patterns and Processes. *PLOS ONE* **8**, e62390 (2013).
321 26. Cook, M. B., McGlynn, K. A., Devesa, S. S., Freedman, N. D. & Anderson, W. F. Sex
322 Disparities in Cancer Mortality and Survival. *Cancer Epidemiol. Biomark. Prev. Publ.*
323 *Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **20**, 1629–1637 (2011).
324 27. Coggon, D., Croft, P., Cullinan, P. & Williams, A. ASSESSMENT OF WORKERS
325 PERSONAL VULNERABILITY TO COVID-19 USING COVID-AGE. *medRxiv*
326 2020.05.21.20108969 (2020) doi:10.1101/2020.05.21.20108969.
327 28. Waters, A.-M., Trinh, L., Chau, T., Bourchier, M. & Moon, L. Latest statistics on
328 cardiovascular disease in Australia. *Clin. Exp. Pharmacol. Physiol.* **40**, 347–356 (2013).
329 29. Nickbakhsh, S. *et al.* Epidemiology of Seasonal Coronaviruses: Establishing the Context
330 for the Emergence of Coronavirus Disease 2019. *J. Infect. Dis.* **222**, 17–25 (2020).

331 30. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the
332 transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**,
333 860–868 (2020).
334 31. Cowling, B. J. *et al.* Increased Risk of Noninfluenza Respiratory Virus Infections
335 Associated With Receipt of Inactivated Influenza Vaccine. *Clin. Infect. Dis.* **54**, 1778–
336 1783 (2012).
337 32. Tsagarakis, N. J. *et al.* Age-related prevalence of common upper respiratory pathogens,
338 based on the application of the FilmArray Respiratory panel in a tertiary hospital in
339 Greece. *Medicine (Baltimore)* **97**, e10903 (2018).
340 33. Ng, K. W. *et al.* Pre-existing and de novo humoral immunity to SARS-CoV-2 in humans.
341 *bioRxiv* 2020.05.14.095414 (2020) doi:10.1101/2020.05.14.095414.
342 34. Grifoni, A. *et al.* Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans
343 with COVID-19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501.e15 (2020).
344 35. BONE MARROW, THYMUS AND BLOOD: CHANGES ACROSS THE LIFESPAN.
345 *Aging Health* **5**, 385–393 (2009).
346 36. Blank, C. U. *et al.* Defining 'T cell exhaustion'. *Nat. Rev. Immunol.* **19**, 665–674 (2019).
347 37. Zheng, M. *et al.* Functional exhaustion of antiviral lymphocytes in COVID-19 patients.
348 *Cell. Mol. Immunol.* 1–3 (2020) doi:10.1038/s41423-020-0402-2.
349 38. De Biasi, S. *et al.* Marked T cell activation, senescence, exhaustion and skewing towards
350 TH17 in patients with COVID-19 pneumonia. *Nat. Commun.* **11**, 3434 (2020).
351 39. Diao, B. *et al.* Reduction and Functional Exhaustion of T Cells in Patients with
352 Coronavirus Disease 2019 (COVID-19). *medRxiv* 2020.02.18.20024364 (2020)
353 doi:10.1101/2020.02.18.20024364.
354 40. Malandro, N. *et al.* Clonal Abundance of Tumor-Specific CD4+ T Cells Potentiates
355 Efficacy and Alters Susceptibility to Exhaustion. *Immunity* **44**, 179–193 (2016).

41. Moore, J. B. & June, C. H. Cytokine release syndrome in severe COVID-19. *Science* **368**, 473–474 (2020).

42. Desdín-Micó, G. *et al.* T cells with dysfunctional mitochondria induce multimorbidity and premature senescence. *Science* **368**, 1371–1376 (2020).

43. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).

44. Funk, S. *https://github.com/sbfnk/socialmixr*. (2020).

45. Stan. *stan-dev.github.io* //mc-stan.org/.

Methods

In the 'Analytical models' section below we include technical details on the deterministic model used to compare theory with data in Figure S3. In the 'Bayesian models' section below we include technical details on the statistical model used to generate the results displayed in Figure 2, S4-S6 and Table S1.

Analytical models

We consider three models: one with age-independent spreading, one with contact-based spreading and no transmission from those infected with mild symptoms and one with contact-based spreading where those with mild symptoms are as contagious as those with severe symptoms. Throughout this paper we use the term 'severe' to correspond to hospitalisations. The first model is a simple model where we assume that transmission and exposure are age-independent and that the probability of subsequent severe infection is proportional to $e^{0.044a}$, where a is age. This model predicts hospitalisation rates to rise as a pure exponential with exponential rate 0.044 years^{-1}.

In our contact based models, we assume that risk of coronavirus infection is of the form

$$P(mild\ infection) = P(exposure)P(mild\ infection\ given\ exposure)$$

$$P(serious\ infection) = P(mild\ infection)P(serious\ infection\ given\ mild\ infection)$$

and that $P(mild\ infection\ given\ exposure)$ is age-independent while $P(serious\ infection\ given\ mild\ infection)$ is proportional to $e^{0.044a}$. We also assume that $P(exposure)$ for someone aged i is proportional to the number of people infected at age j times the amount of contact between age i and age j (as measured by a contact matrix which

we will call C). Then the number of infected people can be modelled as a discrete time Markov process. In what follows, we consider only new cases each time step, which is equivalent to having one time step being the length of time someone is infectious for. We also make the approximation that the number of susceptibles is much larger than the number of infected and recovered/dead.

If individuals with only mild symptoms do not transmit, we can ignore them in our model. If the number of people severely infected at time t, for each age group, is the vector $\boldsymbol{n}_t$ then we have (up to a constant of proportionality setting how fast the epidemic grows)

$$\boldsymbol{n}_{t+1} = E.C.\boldsymbol{n}_t \qquad \text{Eq. 1}$$

where C is the contact matrix and E is a diagonal matrix with $e^{0.044a}$ on the diagonal. Since all elements of C are greater than zero, the Markov chain is strongly connected and therefore, regardless of initial conditions, $\boldsymbol{n}_t$ will be dominated by the term

$$\boldsymbol{n}_t \approx \lambda^t \boldsymbol{v} \qquad \text{Eq. 2}$$

where $\boldsymbol{v}$ is the eigenvector of $E.C$ with the largest eigenvalue, λ. Therefore the normalised age-distribution of new (and cumulative) severe-symptom cases will converge to $\boldsymbol{v}$. The predicted incidence will then be proportional to $\boldsymbol{v}$ divided by the number of people in each age group, which comes from the population age distribution.

If the mild-symptom individuals are as contagious as the severe-symptom individuals, we let the number of infected (both mild and severe) individuals at time t, for each age group, be the vector $\boldsymbol{m}_t$. Then, up to a constant, we have

$$\boldsymbol{m}_{t+1} = \boldsymbol{C}.\boldsymbol{m}_t \qquad \text{Eq. 3}$$

where C is the contact matrix. Again, since the Markov chain is strongly connected, regardless of initial conditions, $\boldsymbol{m}_t$ will be dominated by the term

$$\boldsymbol{m}_t \approx \boldsymbol{\lambda}^t \boldsymbol{v} \qquad \text{Eq. 4}$$

where $\boldsymbol{v}$ is now the eigenvector of C with the largest eigenvalue λ. The age-distribution of severe-symptom cases would then be proportional to $E.\boldsymbol{v}$. The overall constant is then an arbitrary fitting parameter, which we fit to the actual incidence data.

<u>Bayesian models</u>

In this subsection we describe a model focusing on severely infected individuals. This model is the basis of the Bayesian analysis, i.e., where the unknown parameters of the statistical model are fit to data using MCMC methods. In what follows, the unknown, i.e., fitted, parameters are highlighted in bold.

The number of severe infections (hospitalisations) at time *t+1* in age group *i* is denoted by ${Sev^{t+1}}_i$. Similarly, ${Mild^{t+1}}_i$ denotes the number of mild infections (i.e., infections that are not severe). The number of severe infections that arise on a particular day in a given age group has the distribution:

$$Sev^{t+1}{}_i - Sev^t{}_i \sim NegativeBinomial\left(\mu_i^t, (\mu_i{}^t)^\sigma\right) \qquad \text{Eq. 5}$$

where

$$\mu_i^t = \boldsymbol{P_i^{severe}} \mathrm{B}_i^t \ , \qquad \text{Eq. 6}$$

$$S_i^t = H_i - Sev_i^t - Mild_i^t - R_i^t, \qquad \text{Eq. 7}$$

and

$$\mathrm{B}_\mathrm{i}^\mathrm{t} = S_i^t \sum_j \left(\boldsymbol{\beta_{sc}} C_{ij} \frac{Sev_j^{t+\tau}}{H_j} + \boldsymbol{\beta_{mc}} C_{ij} \frac{Mild_j^{t+\tau}}{H_j} + \kappa(\boldsymbol{\beta_{us}} Sev_j^{t+\tau} + \boldsymbol{\beta_{um}} Mild_j^{t+\tau})\right) \qquad \text{Eq.8}$$

In Eq.s 5-8 μ_i^t is the mean number of new severe infections produced in age group i on day t, B_i^t is the force of infection for individuals in age group i on day t that are susceptible to infection (denoted S_i^t). In addition, P_i^{severe} represents the probability of severe disease given infection in age group i, β_{sc} and β_{mc} represent the contact-dependent transmission rates from severe and mildly infected individuals respectively. β_{su} and β_{mu} represent the contact-independent, i.e., environmental transmission rates from severe and mildly infected individuals respectively. The parameter κ is a simple scaling parameter which does not alter the analysis that is included in order to ensure that the contact-dependent and contact-independent transmission rates are on a comparable scale (i.e., to allow fitting of the ratio of these terms; $\kappa = \bar{\rho}/\bar{H}$ where ρ_i is the mean number of contacts that a particular age group, indexed i, makes with other age groups, $\bar{\rho}$ is the mean of $\boldsymbol{\rho}$ and H is the mean of $\boldsymbol{H}$). Finally, τ represents the average delay between the force of infection and the time when a patient is admitted to hospital, which we take to be 0 days with several other values (representing $\tau > 0$) chosen in the sensitivity analysis (see *outline points* below for further details). See Table S2 for a complete set of fitted parameter descriptions. Thus we model the infection process as a negative binomial distribution. This accounts for the variation in count data associated with the occurrence of individual cases in a given age group on a given day, while allowing for potential over-dispersion in count data which may arise, for instance, through the aggregation of French regions, having distinct epidemics. In addition, we model the occurrence of severely infected individuals in an age group given a set number of daily infections in that age group, as a binomial distribution (taken together the binomial and negative binomial distributions result in an overall negative binomial distribution for the occurrence of severe cases, Eq. 5).

The following outline points clarify the choices made in formulating the model:

- In our model, the number of new hospitalisations at time *t* depends on the number of people in hospital at time *t*+τ. Our choice of τ=0 comes from assuming a typical patient would be admitted to hospital ~10 days [43] after infection and stay in hospital

for ~10 days [43]. Therefore hospital admissions at time *t* would depend on prevalence at *t*-10 and prevalence at *t*-10 would lead to hospital occupancy anywhere from *t*-10 to *t*+10. Taking the key points from this interval gives $\tau = 0, \tau = 5$ and $\tau = 10$. We chose to match the simplicity of our approach to the simplicity of our purpose and hence assumed that τ=0 in our main analysis. For comprehensiveness, however, we assumed the other key points from the interval in our sensitivity analyses ($\tau = 5, \tau = 10$).

- Note that the model assumes two components of infection: contact-based and contact-independent infection. Contact-based infection is proportional to social mixing patterns for France recorded in the COMES-F survey (table S3), and is scaled by the prevalence of severe and mild infection in the age groups that have contact with the focal age group. Contact-independent infection is proportional to the absolute number of severe and mild cases in each age group. The contact-independent term reflects the shedding of virus particles into the environment which may then be acquired as aerosolised particles or through contact with infected surfaces (and for this reason is density rather than frequency-dependent).

- The model is fit to data for consecutive epidemic days $t = 1..T$ and for age groups $i = 1..8$ where the age range for each group is ("0-19","20-29","30-39","40-49","50-59","60-69","70-79",">=80") and where the mean age for each age group is $A_i = (9.75, 24.50, 34.58, 44.63, 54.47, 64.42, 73.79, 87.23)$. The French age distribution that was used was taken from the socialmixr dataset in R [44]. Day 1, the first day of the dataset, was 01/03/2020. By default $T = 24$ corresponding to the day that we assumed lockdown effects (which commenced on day 17/03/2020 in France) percolated through to new hospital admissions (i.e., 17 + ~7=~24 with an assumption of a lower bound of 7 days for time from exposure to hospital admission). Note that these factors are varied in the sensitivity analysis (Fig. S6).

- The model calculates the number of mild cases by dividing the number of severe cases (i.e., numbers in hospital) by the probability of severe disease for the respective age groups, i.e., $Mild^{t+1}{}_i = (Sev^{t+1}{}_i / P_i^{severe}) - Sev^{t+1}{}_i$. Note that, as we have confirmed using independent simulation, this is a good assumption as long as the epidemic is growing. However, simulations suggest that contrasting removal rates for different age groups and different infection types (i.e., mild vs severe) can lead to a lack of robustness in this assumption beyond the epidemic peak. Note that since we fit up to lockdown the assumption is valid for our purposes.

Note that the model implementation was performed in RStan [45] using R. The RStan implementations involved 4 chains, 10,000 iterations (of which 1,000 were warm-up), tree depth=15 and adapt_delta=1-(10^{-8}). We obtained the dataset for cumulative age-distributed numbers hospitalised from 'dailyHospCounts_allReg.csv' from [19]. We calculated a dataset of removed (recovered/dead) from the files 'SIVIC_daily_numbers_region_corrected_histo_20200508.csv' and 'SIVIC_total_numbers_region_corrected_histo_20200508.csv' from [19].

Model with deviations

The baseline model indicated that an exponential relationship between probability of severe COVID-19 given infection and age is justified. We therefore extended the baseline model to include an assumption of an exponential form and a single key deviation. This model is the basis of Fig. 2 and Table 2 main text (and is reproduced for comparison in Fig. S5A) , i.e., we allowed each age-cohort to deviate, denoted D_i, in the probability of severe disease given infection, from the exponential relationship, i.e.

$$P_i^{severe} = \frac{P_\eta^{severe} e^{mA_i}}{e^{mA_\eta}} - \boldsymbol{D_i} \qquad \text{Eq. 9}$$

where A_i is the mean age of age group i and A_η is the mean age of the oldest age cohort.

Model with further deviations (sensitivity analysis)

In addition we allowed for a further two deviations. These models are the basis of Fig. S5. In the first of the additional model deviations (Fig. S5B), we allowed each age-cohort to deviate in the probability of infection given exposure, denoted E_i, from a uniform relationship, i.e.

$$\mathrm{B_i^t} = S_i^t \boldsymbol{E_i} \sum_j (\beta_{sc} C_{ij} \frac{Sev_j^{t+\tau}}{H_j} + \beta_{mc} C_{ij} \frac{Mild_j^{t+\tau}}{H_j} + \beta_{us} Sev_j^{t+\tau} + \beta_{um} Mild_j^{t+\tau}) \qquad \text{Eq.10}$$

In the second of the additional model deviations (Fig. S5C), we allowed each age-cohort to deviate in the infectiousness of mildly infected individuals, denoted T_j^m for mildly infected cases, and denoted T_j^s for severely infected cases, from a uniform relationship, i.e.

$$\mathrm{B_i^t} = S_i^t \sum_j (\beta_{sc} C_{ij} \frac{Sev_j^{t+\tau}}{H_j} + \beta_{mc} C_{ij} (1 - \boldsymbol{T_j^m}) \frac{Mild_j^{t+\tau}}{H_j} + \beta_{us} Sev_j^{t+\tau} + \beta_{um} \boldsymbol{T_j^m} Mild_j^{t+\tau}) \qquad \text{Eq. 11}$$

For comprehensiveness, we also considered a variant on the third deviation (Fig. S5D), in which we allowed each age-cohort to deviate in the infectiousness of both mildly and severely infected individuals (denoted T_j^s for severely infected cases), i.e.,

$$\mathrm{B_i^t} = S_i^t \sum_j (\beta_{sc} C_{ij} \frac{Sev_j^{t+\tau}}{H_j} (1 - \boldsymbol{T_j^s}) + \beta_{mc} C_{ij} (1 - \boldsymbol{T_j^m}) \frac{Mild_j^{t+\tau}}{H_j} + \beta_{us} Sev_j^{t+\tau} (1 - \boldsymbol{T_j^s}) + \beta_{um} \boldsymbol{T_j^m} Mild_j^{t+\tau}) \qquad \text{Eq.12}$$

Data sources

A full list of data sources can be found in Table S3.

Out of the countries we have gathered age-stratified data for, some have reported hospitalisation rates, some all confirmed cases and some both. We have included all hospitalisation data we have collected and only excluded data on all confirmed cases for Italy, since that report came from early in the outbreak, when most, but not all, cases were hospitalisations, therefore giving an exponential relationship with a shallow slope.

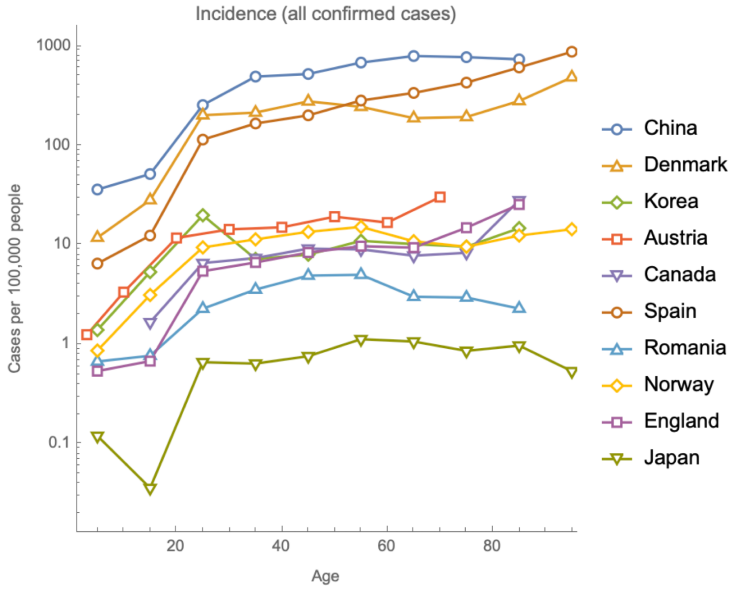
For Fig. 1C we used hospitalisation rates and mortality for each age group with non-zero entries for both males and females. We used data from the following countries: France (hospitalisations), England and Wales (all confirmed cases and mortality), England (CHESS program, hospitalisations) and Spain (all confirmed cases, hospitalisations and mortality).

For Fig. S3 we used pre and post lockdown contact matrices from [19] which are filtered versions of the contact matrices from socialmixr [44].

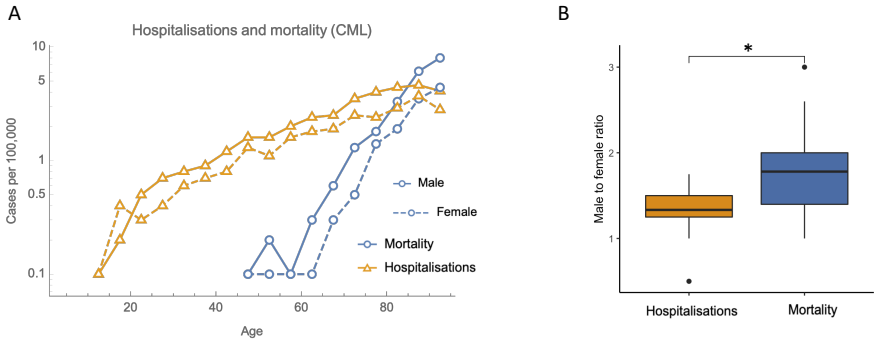
For the rate of thymus decline, we used the estimate from[5], which quotes a half-life of 15.7 years, corresponding to an exponential with rate 0.044 $years^{-1}$ or equivalently a rate of 4.5% per year. We used these same numbers when describing the increase in disease risk with age.

We calculated an R^2 value from the hospitalisation data by focusing on the datasets with the most detailed age-stratification (France, Spain and Denmark) and the age groups with most cases (ages 50-90). Fitting a linear model to the log of the hospitalisation rates gave an exponential rate of 0.046 $years^{-1}$, or equivalently a rate of 4.7% per year with a 95% CI of 4.2-5.2% per year, leading to a 95% CI for the doubling time of 14-17 years. We then fit a linear model with a fixed slope of 0.044 $years^{-1}$, to match the thymic involution timescale, and calculated an R^2 value of 0.98.

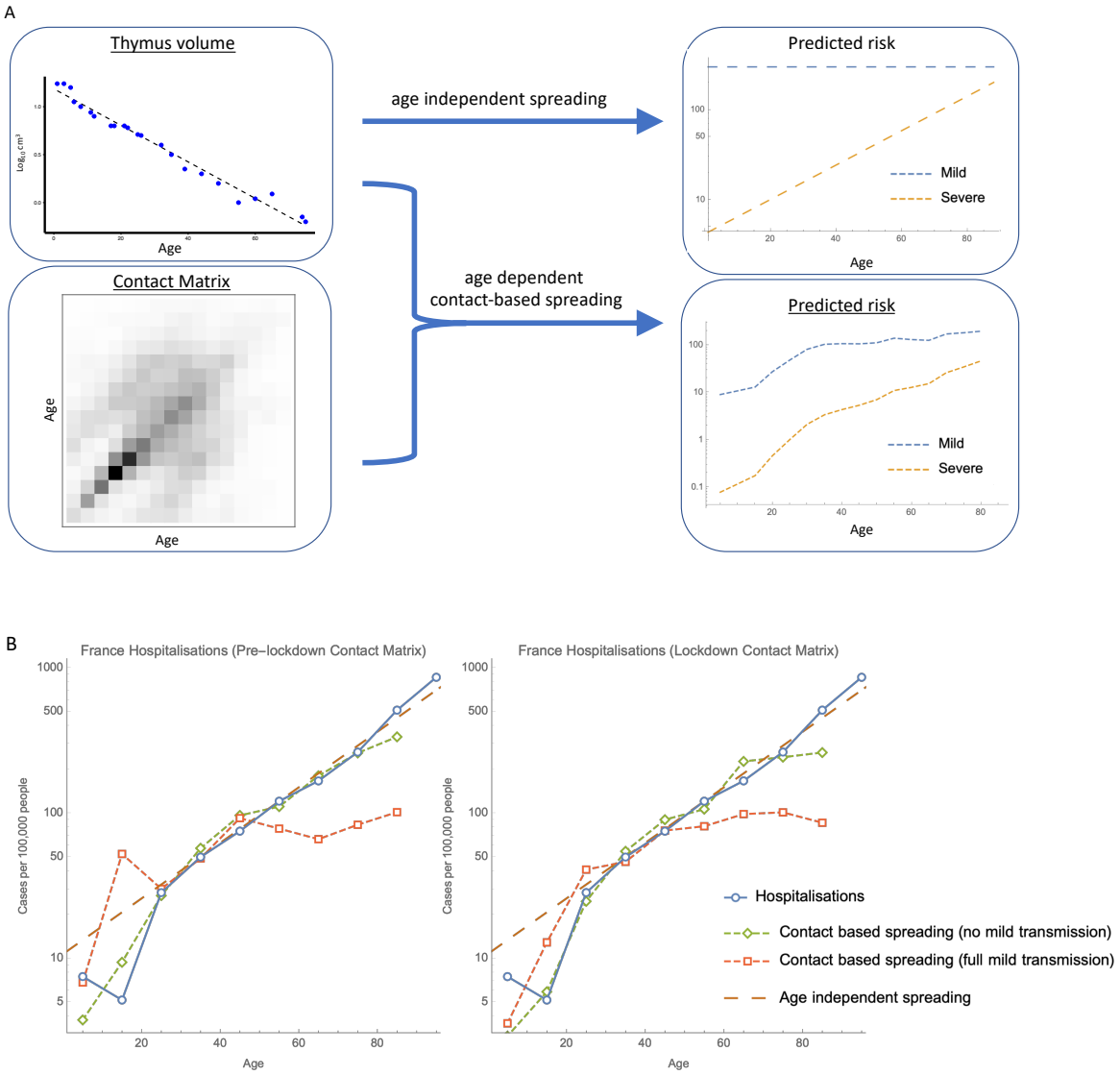
Acknowledgments: We would like to thank Mark Coles, Eamon Gaffney, Jon Chapman, Philip Maini and Marc-Andre Rousseau for comments. In particular we would like to thank Thea Newman for help with an early draft and for the idea to think of risk in terms of effective ages. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data is available in the main text or the supplementary materials or at https://github.com/samIndeed/COVID.



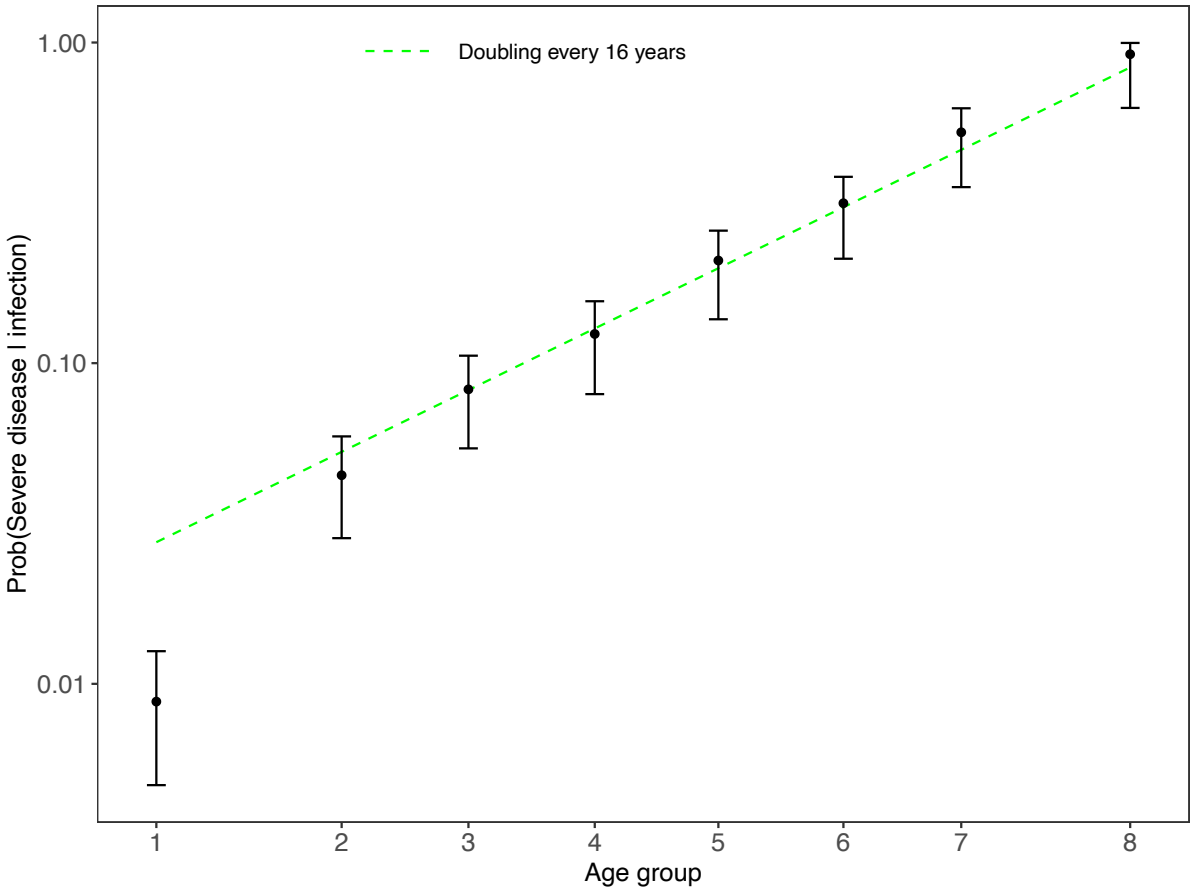
Extended Data Fig. S1. Incidence of all confirmed cases is approximately constant with age for adults and lower in non-adults.



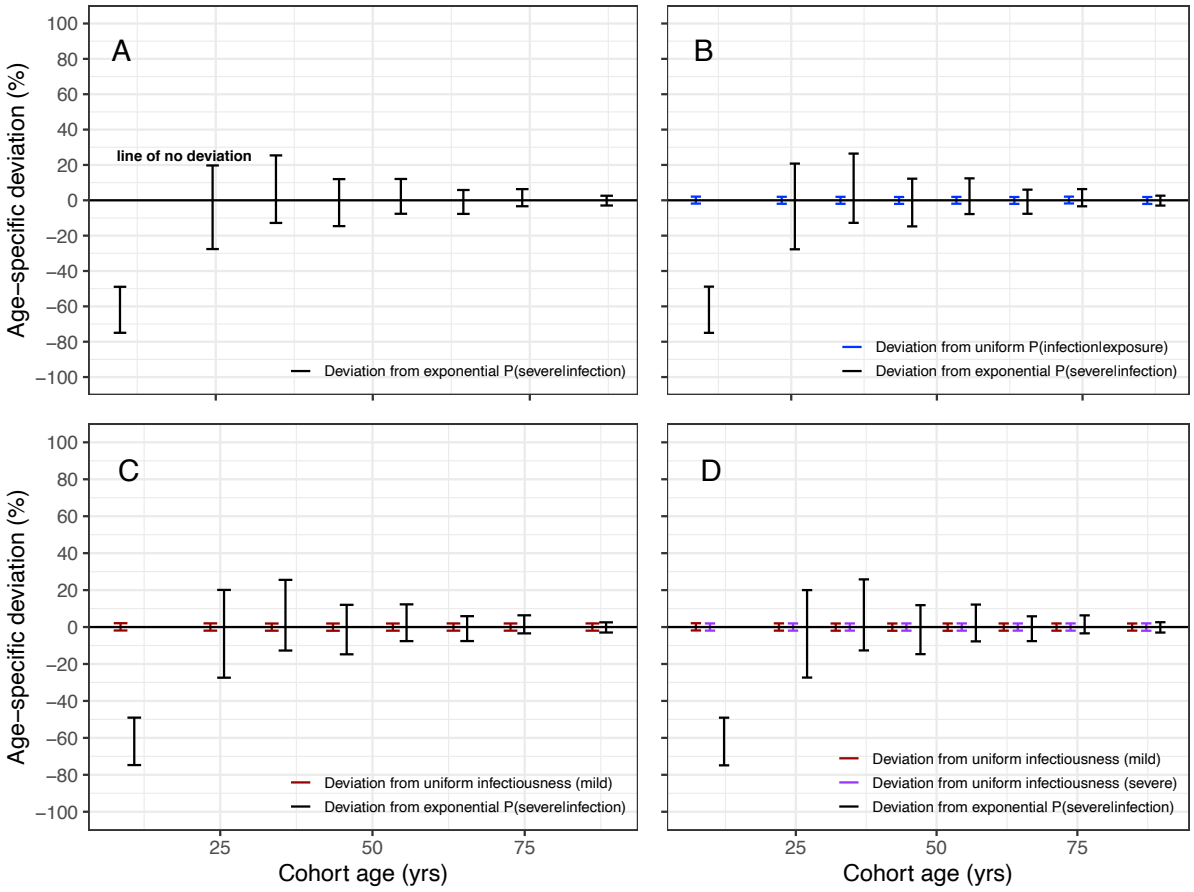
Extended Data Fig. S2. Hospitalisation and mortality rates for CML follow similar trends to COVID-19.



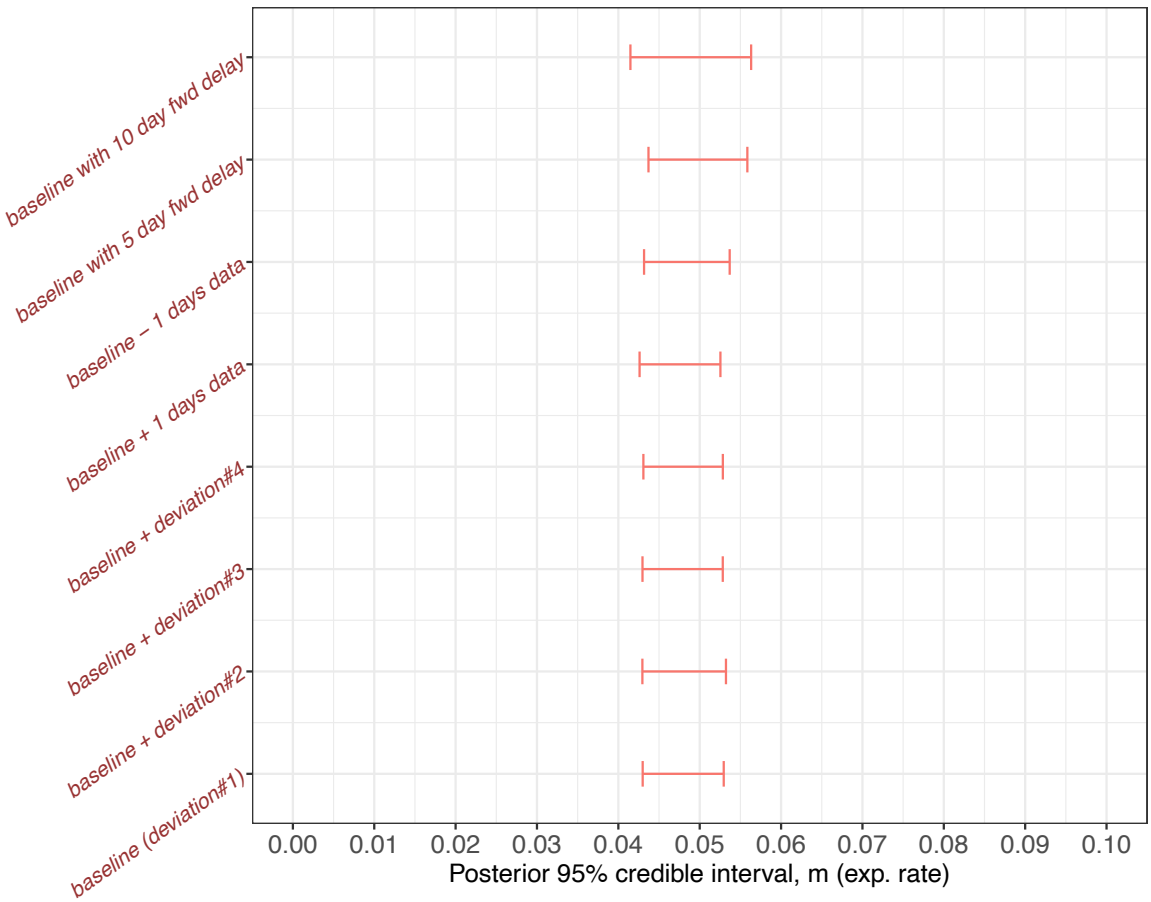
Extended Data Fig. S3. (**A**) Graphical abstract showing the rational for the analytical Markov-chain model (thymus volume data from ref.[5]). If exposure to COVID-19 is age-independent, then a simple model based on immune system declining accurately predicts risk profiles for adults. In this model, the probability of becoming infected is uniform with age and the probability of developing severe symptoms increases exponentially, doubling every 16 years. Alternatively, if the spread of COVID-19 is proportional to contact, as estimated by a contact matrix, then an analytical model can predict similar behaviour for adults and lower risk for non-adults, but only if there is no transmission from mild cases (**B**).



Extended Data Fig. S4. Bayesian analysis of French COVID19 hospitalisation data: the probability that individuals of a given age cohort are hospitalised, given infection, increases exponentially with age. When the probability of hospitalisation given infection was allowed to vary independently for each age cohort, the posterior probabilities of hospitalisation (95% credible intervals in black, with filled circles for mean posterior values) were found to increase at a rate corresponding to thymus degradation for cohorts over the age of 20 (green dashed line) (black dots).



Extended Data Fig. S5. Low COVID-19 hospitalisation among juveniles is associated with additional intrinsic defence beyond strong thymus function instead of age-specific differences in several alternative processes. The likelihood of severe disease increases exponentially with age at a rate matched by thymus degradation, but juveniles deviated significantly from this pattern (A). None of the deviations for the following processes (which may also be explanations for low juvenile severe disease) were significant: age-specific deviations in the probability of infection given exposure (B, blue intervals), or in the infectiousness of mildly infected cases (C, red intervals), or in the infectiousness of mildly and severely infected cases (D, red and purple intervals respectively). The line depicting zero deviation (black, bold, horizontal) passes through each of the credible intervals with the sole exception of the deviation from exponential probability of severe disease for the juvenile age group.



Extended Data Fig. S6. Sensitivity analysis to test the assumptions of the Bayesian analysis of French COVID19 hospitalisation data. Bayesian 95% credible intervals for the rate parameter in the exponential dependence of probability of severe disease. Baseline represents the model in Fig. 2. Baseline+deviation#2 represents the additional inclusion of age-specific deviations from uniform probability of infection given exposure. Baseline+deviation#3 represents the additional inclusion of age-specific deviations from uniform infectiousness of mild cases. Baseline+deviation#4 represents the additional inclusion of age-specific deviations from uniform infectiousness of mild cases and severe cases. Baseline+1 days data represents the extension of the timeseries fit to the model by 1 further day (i.e. baseline model run with 25 days data). Baseline-2 days data represents the truncation of the timeseries fit to the model by 2 days (i.e. baseline model run with 22 days data). Baseline with 5 day fwd delay (and similarly with 10 day fwd delay) represents the baseline model with the rate of occurrence of hospital admissions at time t depending on the numbers hospitalised at time $t + 5$ (and similarly depending on the numbers hospitalised at time $t + 10$).

Extended Data Table S1. Bayesian 95% credible intervals for epidemiological parameters from French COVID19 hospitalisation data.

Epidemiological parameter	***Mean***	***2.5% (lr 95% threshold)***	***97.5% (upr 95% threshold)***
$\boldsymbol{\beta_{sc}/\beta_{mc}}$	14.89	0.04	45.11
$\boldsymbol{\beta_{su}/\beta_{mu}}$	0.38	0.01	1.60
$\boldsymbol{\beta_{sc}/\beta_{su}}$	9.28	0.02	38.84
$\boldsymbol{\beta_{mc}/\beta_{mu}}$	0.256	0.007	0.988
$\boldsymbol{P^{severe}_{>80y}}$	0.781	0.391	0.992
Additional juvenile protection	63.4%	48.9%	*75%*

Extended Data Table S2. Parameter definitions, and prior distributions, for the Bayesian analysis. All prior distributions were chosen to be non-informative. Posterior parameter distributions were obtained using Hamiltonian Monte Carlo, RStan version 2.19.3 [45], R version 3.6.3.

	Parameters to be fitted	
m	Probability of severe disease given infection, age coefficient	$\sim$ *gamma(1,20)*
β_{sc}	Contact-based transmission rate from severely infecteds	$\sim$ *gamma(1,20)*
β_{mc}	Contact-based transmission rate from mildly infecteds	$\sim$ *gamma(1,20)*
β_{su}	Contact independent transmission rate from severely infecteds	$\sim$ *gamma(1,20)*
β_{mu}	Contact independent transmission rate from mildly infecteds	$\sim$ *gamma(1,20)*
$(P_i^{severe}$	Inverse of probability of severe disease given infection, age group *i*	$\sim$ *cauchy(0, 1)*
σ^{-1}	Inverse of negative binomial dispersion exponent	$\sim$ *normal(2,0.5)*
	Parameters to be fitted, model with deviations	
P_8^{severe}	Prob. of severe disease given infection, for oldest age group $i = 8$	$\sim$ *cauchy(0, 1)*
D_i	Deviation from exponential prob. of severe disease given infection, age group *i*	$\sim$ *normal(0,0.01)*
E_i	Deviation from uniform prob. of infection given exposure, age group *i*	$\sim$ *normal(0,0.01)*
T_j	Deviation from uniform infectiousness of infecteds, age group *i*	$\sim$ *normal(0,0.01)*
	Data variables	
H_i	Population size, age group *i*, *France*	*pre-existing data*
Sev_i^t	#Hospitalised COVID-19 patients, age group *i*, *France*	*pre-existing data*
$Mild_i^t$	#Infected with COVID-19 and not hospitalised, age group *i*, *France*	*Calculated from parameter fits*
R_i^t	#Removed, i.e. dead or released, age group *I*, *France*	*pre-existing data*
C_{ij}	#Contacts with age group *j,* age group *i, France*	*pre-existing data*

Extended Data Table S3. List of data sources for COVID-19 hospitalisations, cardiovascular disease and CML.

France (daily data)	https://zenodo.org/record/3889894
France (cumulative)	https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/?fbclid=IwAR0JUQuhfHhu8OQsX7xo3FDgl-cWYiiDO7z-oCPnANuXUQIax1WqZyTr8FY
Spain	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/Actualizacion_70_COVID-19.pdf
Denmark	https://files.ssi.dk/COVID19-overvaagningsrapport-08042020-zm92
USA (accessed April 2020)	https://gis.cdc.gov/grasp/COVIDNet/COVID19_3.html
Canada (accessed June 2020)	https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html
England (CHESS program)	https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/880925/COVID19_Epidemiological_Summary_w17.pdf
England and Wales	https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales
Switzerland (accessed April 2020)	https://www.bag.admin.ch/dam/bag/de/dokumente/mt/k-und-i/aktuelle-ausbrueche-pandemien/2019-nCoV/covid-19-datengrundlage-lagebericht.xlsx.download.xlsx/200325_Datengrundlage_Grafiken_COVID-19-Bericht.xlsx
China	http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51
South Korea (accessed April 2020)	http://ncov.mohw.go.kr/bdBoardList_Real.do?brdId=1&brdGubun=11&ncvContSeq=&contSeq=&board_id=&gubun=
Austria (accessed April 2020)	https://info.gesundheitsministerium.at/
Romania (accessed April 2020)	https://datelazi.ro/
Norway	https://www.fhi.no/contentassets/e110607a67df46cbba8e30a443264a73/vedlegg/tidligere-dagsrapporter/2020.06.10-dagsrapport-norge-covid-19.pdf
Japan (accessed April 2020)	https://toyokeizai.net/sp/visual/tko/covid19/
Australia - Cardiovascular disease	https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/cardiovascular-health-compendium/data
UK - CML	https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-cml