PLOS ONE

RESEARCH ARTICLE

Causal, Bayesian, & non-parametric modeling of the SARS-CoV-2 viral load distribution vs. patient's age

Matteo Guardiani[1,2]*, **Philipp Frank**[1,2], **Andrija Kostić**[1,2], **Gordian Edenhofer**[1,2], **Jakob Roth**[1,2], **Berit Uhlmann**[4], **Torsten Enßlin**[1,2,3]

1 Max Planck Institute for Astrophysics, Garching, Germany, **2** Fakultät für Physik, Ludwig-Maximilians-Universität München, Munich, Germany, **3** Excellence Cluster ORIGINS, Garching, Germany, **4** Süddeutsche Zeitung, Munich, Germany

* matteani@mpa-garching.mpg.de

Check for updates

OPEN ACCESS

Citation: Guardiani M, Frank P, Kostić A, Edenhofer G, Roth J, Uhlmann B, et al. (2022) Causal, Bayesian, & non-parametric modeling of the SARS-CoV-2 viral load distribution vs. patient's age. PLoS ONE 17(10): e0275011. https://doi.org/10.1371/journal.pone.0275011

Editor: Christoph Strumann, Universitatsklinikum Schleswig Holstein Campus Lubeck, GERMANY

Received: July 14, 2021

Accepted: September 9, 2022

Published: October 5, 2022

Copyright: © 2022 Guardiani et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data files are available at https://figshare.com/articles/dataset/Cobas_Dataset_csv/19690864 and https://figshare.com/articles/dataset/LC_480_Dataset_csv/19690846.

Funding: This research was partly supported by the German Aerospace Center and Federal Ministry of Education and Research through the project Universal Bayesian Imaging Kit – Information Field Theory for Space Instrumentation (Förderkennzeichen 50OO2103).

Abstract

The viral load of patients infected with SARS-CoV-2 varies on logarithmic scales and possibly with age. Controversial claims have been made in the literature regarding whether the viral load distribution actually depends on the age of the patients. Such a dependence would have implications for the COVID-19 spreading mechanism, the age-dependent immune system reaction, and thus for policymaking. We hereby develop a method to analyze viral-load distribution data as a function of the patients' age within a flexible, non-parametric, hierarchical, Bayesian, and causal model. The causal nature of the developed reconstruction additionally allows to test for bias in the data. This could be due to, e.g., bias in patient-testing and data collection or systematic errors in the measurement of the viral load. We perform these tests by calculating the Bayesian evidence for each implied possible causal direction. The possibility of testing for bias in data collection and identifying causal directions can be very useful in other contexts as well. For this reason we make our model freely available. When applied to publicly available age and SARS-CoV-2 viral load data, we find a statistically significant increase in the viral load with age, but only for one of the two analyzed datasets. If we consider this dataset, and based on the current understanding of viral load's impact on patients' infectivity, we expect a non-negligible difference in the infectivity of different age groups. This difference is nonetheless too small to justify considering any age group as noninfectious.

Introduction

Children do not seem to be major drivers in the transmission of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in the general population [1]. However, the exact degree to which children and adolescents get infected by, and are able to transmit the virus is not yet well known. Their role in the community spread depends on their susceptibility, symptoms, viral load, social contact patterns, behavior, and existing mitigation strategies as schools and

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

daycares closings. Among all these variables, the viral load plays a fundamental role. The viral load might help to predict disease severity [2] and mortality [3–5] and can serve as a proxy for the infectivity of the patient [6–8]. The severity of a disease, its infectivity, and its mortality are certainly fundamental parameters that must be considered when deciding on best-practice preventative measures to fight the pandemic spread. Research in this direction can enable truly data-driven policymaking like, for example, school openings and focused lockdowns. For this scope, it is important to understand how the viral load depends on the patients' age.

In this work, we examine viral load as a proxy for infectivity. We reanalyze the age-stratified viral load data from Jones et al. [9] in order to better understand the actual relation between these variables. We do this in the hope of gaining insight into fundamental differences reported in the literature regarding the relationship between viral load and age. We achieve this goal by developing a flexible, non-parametric, causal, and Bayesian model to reconstruct the conditional probability density function (PDF) of the viral load given the patient's age. The developed method is a second central result of this work: it can be applied to future studies on SARS-CoV-2, to similar data from other diseases, and also to many causally connected quantities in very different contexts.

The non-parametric PDF reconstruction is regularized by mild assumptions on the smoothness of the underlying statistical processes. In particular, we assume that the log-densities are Gaussian-process realizations, drawn with an a priori flexible correlation kernel parametrized by a Matérn family correlation function. The parameters of this correlation function are then inferred along with the PDF through a variational inference algorithm. To achieve this, we adapt methods developed for information field theory, the information theory for fields [10, 11]. In this context, fields are understood as spatially varying (physical) quantities. The reconstructed PDFs are regarded as scalar fields whose values are defined at each point of the two-dimensional space spanned by age and viral load and represent the probability of observing a given combination of age and viral load. The toolkit of information field theory has proven itself to be successful in a wide range of applications, ranging from 3D tomography [12], over time-resolved astronomical imaging [13], to causality inference [14].

Outline

The rest of this work is structured as follows: in Sec. Related work, we discuss the state of the art of the research on the assessment of the viral load-vs-age dependence. In Sec. Model design, we motivate the need for a causal description and show how this description is built into the model and the inference scheme we adopt. In Sec. Data, we describe how the data has been acquired and processed. We then outline the main results of our study in Sec. Results, focusing on their impact on the infectivity of SARS-CoV-2. Finally, in Sec. Conclusions we summarize the benefits of our approach while highlighting its potential limitations and identifying possible future work directions.

Related work

Since the outbreak of SARS-CoV-2, efforts have been made in order to understand whether certain age groups are more susceptible than others. This could either mean that people from such age groups are more likely to get infected or that they show more severe symptoms compared to older or younger individuals. In addition to this, patients from specific age groups could be more infectious than others, hence more likely vehiculating the disease. To shed light on these problems, viral load—which is a proxy for infectivity—can be a useful tool. It is measured by reverse transcription PCR (RT-qPCR) assays from nasopharyngeal and oropharyngeal swabs via the so-called cycle threshold (Ct) value. The viral load is the virus concentration

in the upper respiratory tract and it is usually expressed as the number of viral RNA copies per mL of sample or entire swab specimen or simply by the Ct value.

Several works have analyzed whether viral loads differ between children and adults [1, 15–27], and between young children and adolescents [28–30], and have led to conflicting results. At least eight studies from different countries have concluded that SARS-CoV-2 viral RNA loads among children and adults were comparable [1, 15, 16, 19, 21–24]. In these studies, dependence between age and viral load has been tested using one and two-way analysis of variance (ANOVA) [15] and median of the viral load. A further study [17] found that mean and median viral load values did not vary conspicuously by age but noted that the highest values were measured in patients born from 1995 to 2009. In contrast, at least five studies reported significant differences in viral loads of young children and adults [18–20, 25–27]. Euser et al. [18] suggested an approximately 16-fold higher viral load in the oldest age group (>79 years), compared to the youngest age group (<12 years). Here, age-group differences in the viral load distribution are assessed making use of the Kruskal-Wallis test and linear regression. Another work [20] estimated the amount of SARS-CoV-2 in the upper respiratory tract of young children (<5 years) to be 10-fold to 100-fold greater than in adults whereas a work from Zachariah et al. [30] showed that mean viral load was significantly higher in infants (<1 year) as compared to older children and adolescents.

Finally, one of the largest and most widely followed studies on the subject—even though the number of children and adolescents included is fairly small—was carried out by Jones et al. [9] in early 2020. Dependence between viral load and age has been tested for different age groups both as categorical data and treating age as a continuous variable. In order to compare the viral load of different age categories, the categorical data has been analyzed in a parametric (Welch's T-test), non-parametric (Mann-Whitney rank test), and Bayesian fashion (modeling viral loads as a mixture of gamma distributions). When considering age as a continuous variable, viral loads have been predicted from age, type of PCR system, and age-PCR system interaction. This study from Jones et al. [9] did not reveal large differences in the viral loads of different age groups, a result that was publicly debated in Germany for its possible implications for school opening policies. In the following, we reanalyze this data. In a very recent publication, Jones et al. [31] extended their initial version of the study. In this newer version, they make use of thin-plate spline regression to conclude that children and adolescents have a slightly lower viral load than adults, but that this difference is unlikely to be clinically relevant.

None of the studies in the literature utilizes the causal framework. In this work, we choose to adopt this framework since it naturally allows to answer the central question of whether the age of a patient causes its viral load and infectivity. It also leads to additional advantages that we will present in the following sections. The problem of deducing causal directions (in particular for the bivariate $x \to y$, $y \to x$ and $x \perp y$ case) from observational data coming from a joint distribution has been introduced by the works of J. Pearl [32] and P. Spirtes et al. [33], further developed by Mooij, J. et al. [34] and is now a central and non-trivial problem in data analysis. For additional details and motivations behind causal inference theory and techniques we refer to their works.

Model design

In Sec. Related work we described the state of the art of statistical analyses performed in order to investigate the age dependence of the viral load. These analyses mostly rely on variance tests or correlation assessments. In the following, we motivate the need for a causal model of the viral load distribution. In fact, in order to explain the age and viral load data collected by Jones et al. [9], a well-behaved model should incorporate basic knowledge about the causal relation

between age and viral load. It should furthermore allow questioning whether the viral load distribution depends on the patients' age and quantify the strength of such dependence, if it exists. We consequently develop a non-parametric causal model and apply it to data.

Motivation

In order to make statements about the factors that contribute to the spread of a disease, we need to rely on data to ground these claims. More importantly, we need to find a model that is capable of identifying and explaining the relationships that underlie the data. Indeed, this should be a minimal requirement for any data-analysis task. The choice of an incorrect model can lead to wrong or misleading conclusions. This is clearly an issue regardless the nature of the data at hand. However, in the case of data describing the pandemic spread the choice of an incorrect model can lead to ineffective or even potentially harmful decisions. It is therefore of vital importance to be able to capture—at least up to a certain level of uncertainty—the interdependences underlying the data.

As discussed in the previous section Sec. Related work, the problem of identifying the relation between age and viral load has been tackled making use of many different techniques. While categorical data analysis and linear correlation analysis struggle to pick up all possible dependences, non-parametric approaches have the drawback of possibly being too (or too little) complicated to account for the actual functional dependence that is inherent to the data. For this reason we want to build a model that is flexible in the sense that it can simply reconstruct independent densities with few degrees of freedom while also being capable of inferring more complex distributions, when needed to describe the data. Moreover, we would like the model to automatically be able to adjust its complexity. This way, the choice of the model is independent from the data analyst's choices and the results are more consistent and reproducible.

This is another benefit of adopting the causal framework. The concepts of "simple" independence and "more complicated" dependence between variables are natural to this framework and are described by the causal graph and the structural causal model of choice [34]. Furthermore, claiming causal dependence is much stronger than simple linear correlation, which also can arise from a confounder or by mere coincidence. Causal dependence is directional. This is a particularly interesting feature as it allows to test for bias in the data-collection process when a unnatural causal direction is detected from the data analysis, as we will discuss in Sec. Results.

Causal structure

The analyzed dataset $d = \{(i, x_i, y_i)\}_{i=1}^{N}$ consists of indexed pairs of age x_i = age$_i$/years and log-viral load $y_i = \log_{10}$(viral RNA copies$_i$/ml) values for each of the N infected patients, extracted from Fig. 6 in [9]. The dataset is shown in Fig 1. For simplicity, we refer to y as the "viral load", where 10^y are the number of viral RNA copies per milliliter of sample or entire swab specimen. We assume that the data points are Poisson process counts drawn from an underlying stationary density distribution $\varrho(x, y)$. Using the causal model shown in Fig 2 we reconstruct the density $\varrho(x, y)$. Our final density reconstructions are displayed in Figs 3 and 4. To model this density, we express it in terms of the underlying age distribution $\varrho(x)$ of the patients times the conditional PDF $p(y|x)$ of the viral load y given the patient's age x,

$$\varrho(x, y) = \varrho(x)\ p(y|x). \tag{1}$$

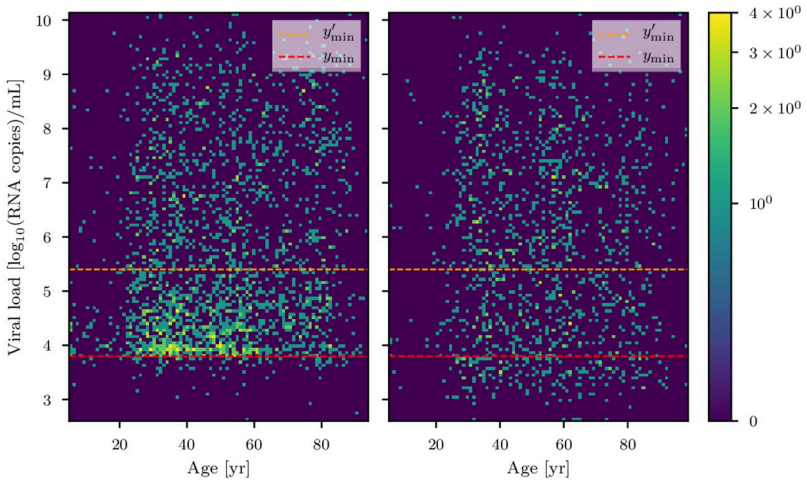

Fig 1. The cobas (left) and the LC 480 (right) datasets. The lower thresholds $y_{\min}$ and $y'_{\min}$ with which the data has been filtered are shown in red and orange, respectively. The number of data counts is color coded with a logarithmic color scheme (see colorbar).

https://doi.org/10.1371/journal.pone.0275011.g001

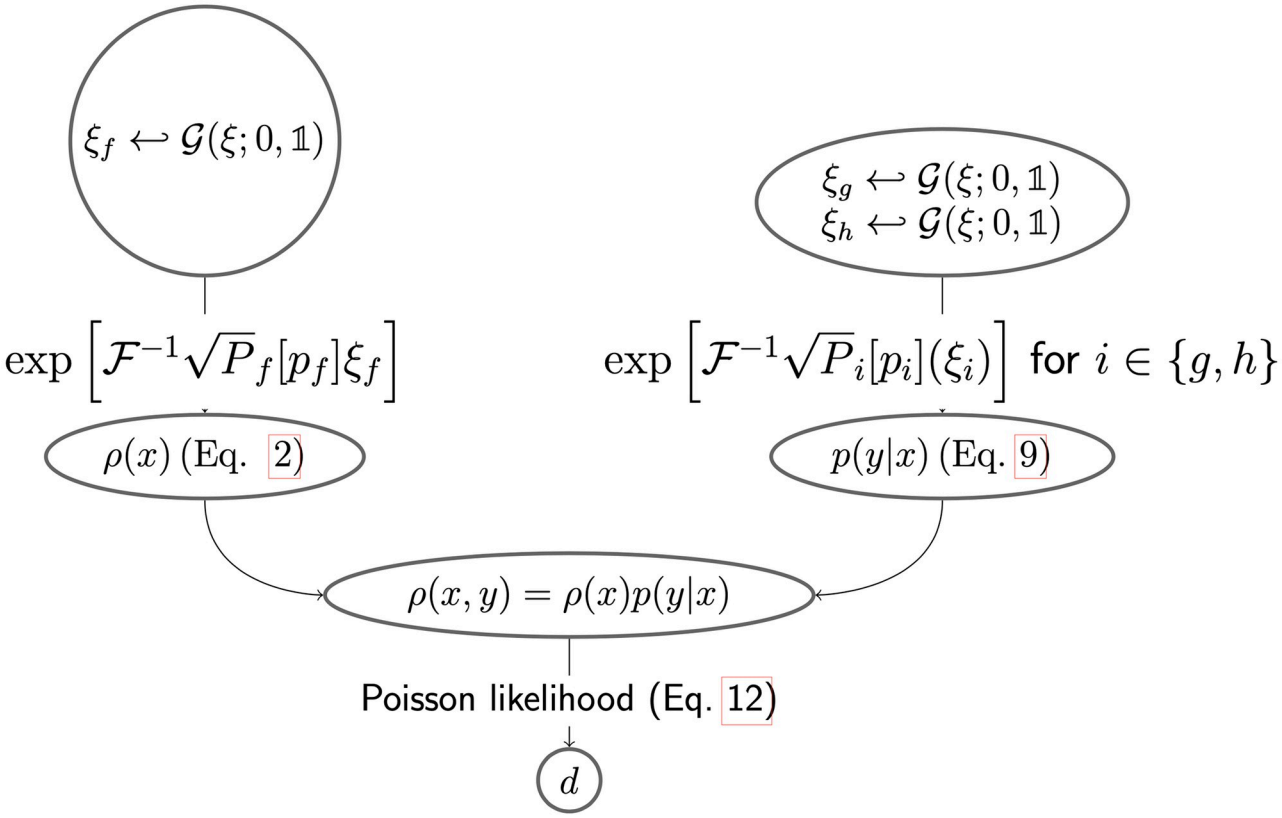

Fig 2. Graph structure of the causal model of age x and viral load y. $\mathcal{F}$ denotes the Fourier transform operator. Starting from the top, standard-normally distributed excitations ξ are drawn from the latent priors for f, g, and h. They are then transformed into the non-parametric signal $\rho(x, y)$ (Eq 9) by taking the inverse Fourier transform of the Matérn kernel parametrization described in Eq 5. Finally, the signal $\rho(x, y)$ is compared to the data d through a Poissonian likelihood (Eq 12).

https://doi.org/10.1371/journal.pone.0275011.g002

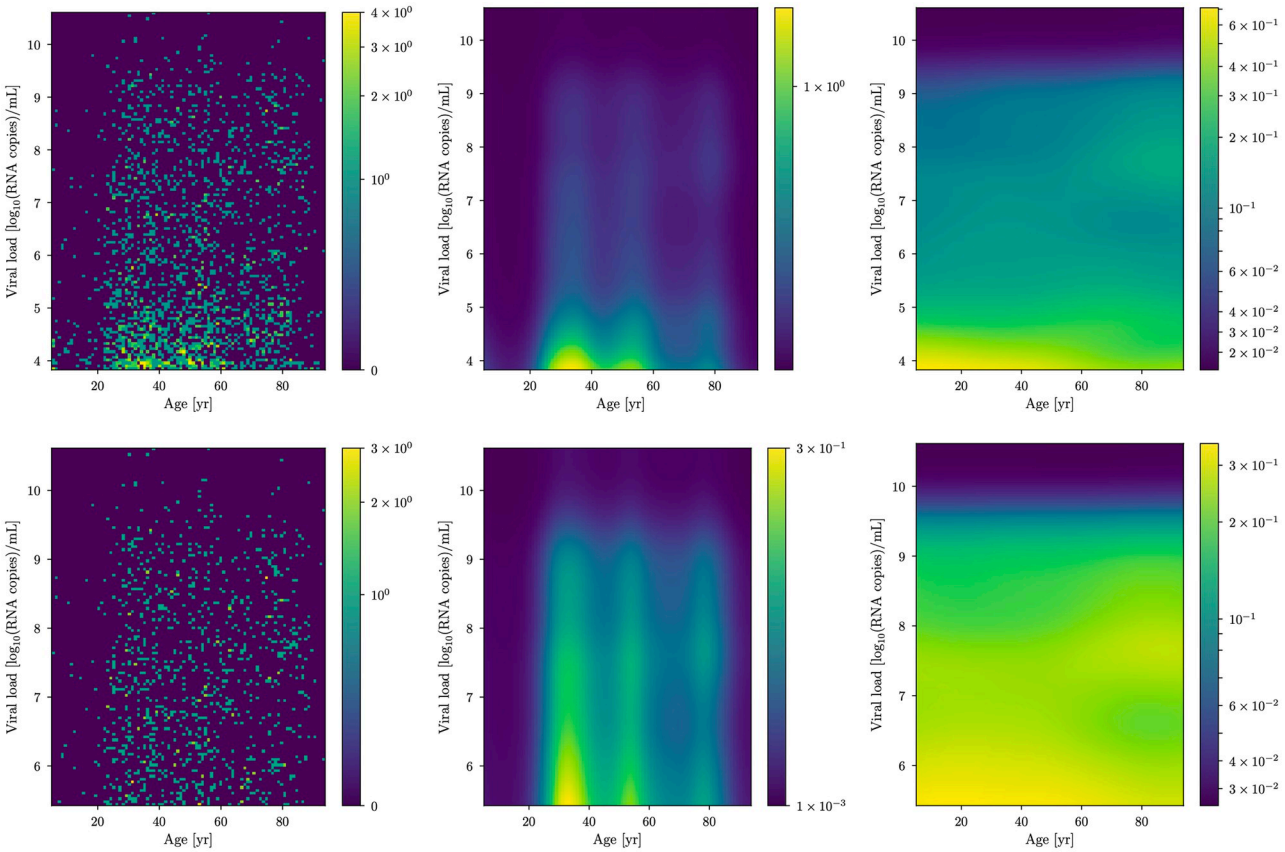

Fig 3. Cobas dataset analysis. Left: The cobas dataset. Middle: the reconstructed density distribution $\varrho(x, y)$ as a function of the age (x) and the viral load (y) in a logarithmic coloring scheme. Right: The 2D conditional probability distribution $p(y|x)$ of the viral load (Eq 9) obtained by fitting the model. The data and the results of the analysis are shown for two different data-filtering thresholds $y_{\min}$ (top) and $y'_{\min}$ (bottom).

https://doi.org/10.1371/journal.pone.0275011.g003

In Eq 1 the causal direction $x \to y$ (age influences viral load) is implicitly introduced. Even though $x \to y$ might appear to be the most intuitive causal direction, since the immune system reaction depends on age and thus age should affect the viral load, we note that selection effects could introduce different apparent causal structures in the data. If the data had been collected in such a way that the viral load (y) was the deciding factor for whether a patient would enter the data sample, with an age (x) dependent threshold, the viral load would impact the age distribution in the sample, leading to an apparent $y \to x$ causal structure.

As a practical example, one could imagine physicians to request a PCR test for a child only if their symptoms were worse compared to those of an adult for which they would have requested a PCR test. Since this and similar selection effects cannot be fully excluded for the analyzed data (see discussion in [9]), we will calculate the Bayesian evidence for the possible causal relations $x \to y$, $y \to x$, and $x \perp y$ (i.e. x and y are independent, $p(y|x) = p(y)$). As a final remark, we note that other effects such as, e.g., delays in data reporting, could in principle also be confounding the simple $x \leftrightarrows y$ causal direction. Li and White [35] have for instance shown how reporting delays can have an effect on the pandemic-spread models. Effectively, a confounding variable z would influence both x and y, namely $x \leftarrow z \to y$. Although we cannot completely exclude the presence of such a confounding variable, the randomization tests that we describe in Subsec. Outline support our implicit assumption of absence of any leading confounding variable.

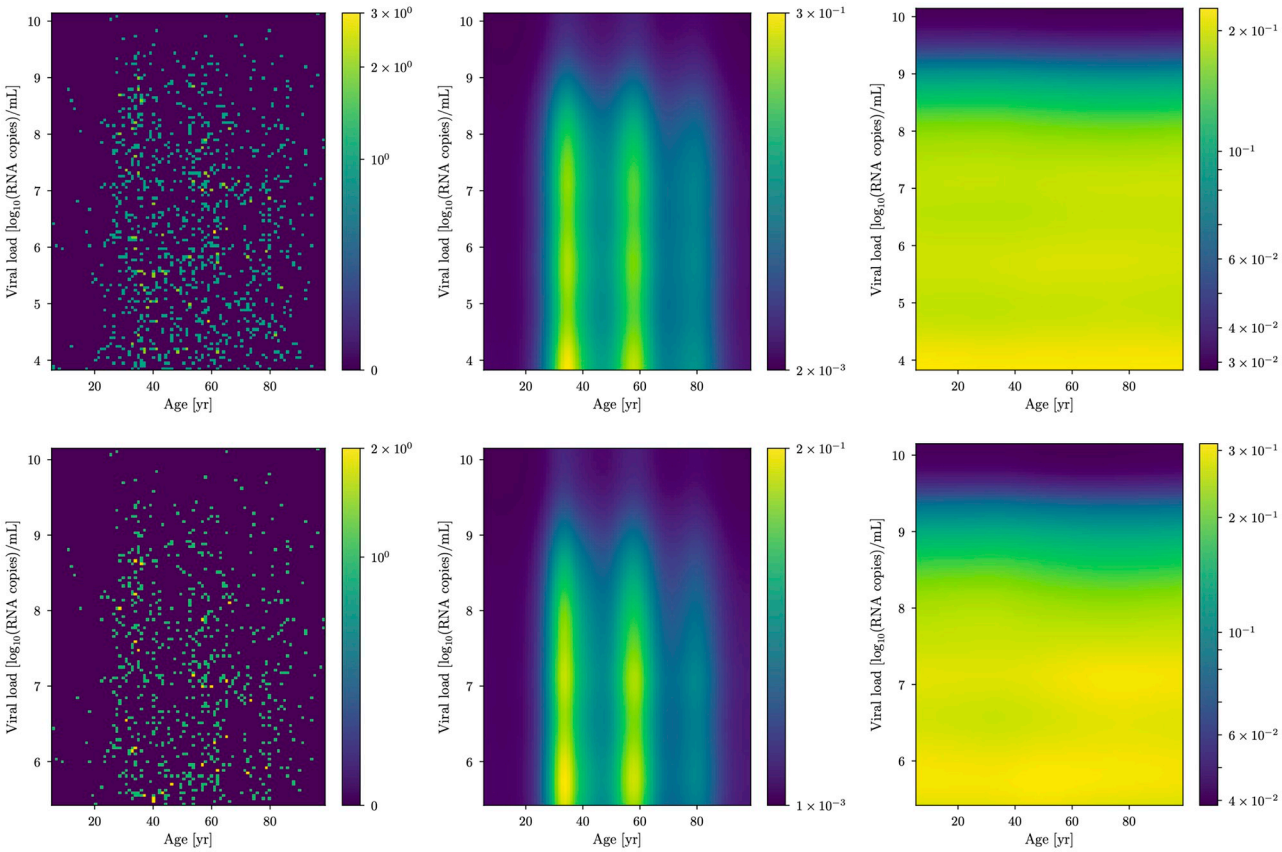

Fig 4. LC 480 dataset analysis. The same as in Fig 3, but for the LC 480 dataset.

https://doi.org/10.1371/journal.pone.0275011.g004

To model the $x \to y$ causal direction, we need to model the age distribution $\varrho(x)$ of the infected patients according to the causal structure introduced in Eq 1. Lacking knowledge on the exact details of the patient selection process, we assume $\varrho(x)$ to be a log-normally distributed random variable. The log-normal distribution is a natural choice since an age density is by definition a strictly positive and continuous quantity. Another natural assumption is the absence of abrupt changes, since no sharp age-selecting processes are expected to have shaped it. We fulfill these assumptions with the choice

$$\varrho(x) = \varrho_0 e^{f(x)}, \tag{2}$$

where $\varrho\, 0 = N/100$ is a reference density and $f : [0, 100] \mapsto \mathbb{R}$ a smooth function centered around zero. We accordingly assume f to be drawn from a zero centered Gaussian process with a prior covariance F

$$\mathcal{P}(f) = \mathcal{G}(f, F) := \frac{1}{\sqrt{2\pi F}} \exp\left(-\frac{1}{2} f^{\dagger} F^{-1} f\right). \tag{3}$$

The covariance

$$F_{xx'} = \langle f(x) f(x') \rangle_{(f)} := \int \mathcal{D}f \;\; \mathcal{P}(f) f(x) f(x') \tag{4}$$

determines the degree of smoothness of the logarithmic distribution function, as well as the

characteristic length scale and the amplitude of its variations. We assume this correlation structure to be translation invariant $F_{xx'} = F(x - x')$, since we only expect it to depend on age differences—and not on a particular age value—and parametrize it with a Matérn kernel. Invoking the Wiener-Khinchin theorem, we can represent such a translation-invariant correlation function in Fourier space with a spectral density of

$$P_f(k) = \frac{a_f^2}{\left[1 + \left(k/k_f\right)^2\right]^{\gamma_f/2}}, \tag{5}$$

with a_f specifying the amplitude of the variations in f, $1/k_f$ the characteristic length-scale above which the variations become uncorrelated, and γ_f the spectral index, which determines the smoothness of the variations. We infer all three covariance parameters $p_f := (a_f, k_f, \gamma_f)$ from the data. In order to ensure that the model is flexible enough to fit the data, we set mildly informative priors on the covariance parameters. We denote by $\mathcal{P}(f|p_f)$ the probability of a specific realization of f given the Matérn kernel parameters p_f, as described by Eqs 3 to 5.

Next, we have to specify the distribution of the viral load given the age, $p(y|x)$. To do so, we note that we can directly model an independent distribution for which $p(y|x) = p(y)$ in the same way as we modeled $\rho(x)$, i.e. by choosing $p(y|x) \propto e^{g(y)}$. Of course, $p(y|x)$ can in general exhibit an arbitrary complicated dependence on x. We model any additional complicated entanglement between the age x and the viral load y with a new function $h(x, y)$. By doing so, we introduce a degeneracy between $g(y)$ and $h(x, y)$, since they can both model y-only dependent structures. In principle, the function $h(x, y)$ can in fact model any function $p(y|x)$ without the necessity of introducing $g(y)$. To solve this problem, we choose

$$p(y|x) \propto \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}}. \tag{6}$$

To ensure that only $g(y)$ models strictly y-dependent features and that all possibly complicated x–dependent features are captured by $h(x, y)$, we have to prevent $h(x, y)$ from modeling any strictly y-dependent structure that has been already captured by $g(y)$. We do this with the denominator $\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}$ in Eq 6, that removes from $p(y|x)$ any y-only structure contained in $h(x, y)$, hence eliminating the degeneracy between $g(y)$ and $h(x, y)$.

We can verify that $h(x, y)$ indeed satisfies the desired property of not encoding any structure that could be represented by $g(y)$ by simply substituting $h(x, y) \mapsto h'(x, y) := h(x, y) + g'(y)$ in Eq 6, where $g'(y)$ depends only on y. This results in h and h' leading to the same conditional PDF

$$\begin{aligned} p'(y|x) \quad &\propto \frac{e^{g(y)+h'(x,y)}}{\int e^{h'(\tilde{x},y)}\mathrm{d}\tilde{x}} = \frac{e^{g(y)+h(x,y)+\cancel{g'(y)}}}{\cancel{e^{g'(y)}}\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}} \\ &= \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}} \propto p(y|x) \end{aligned} \tag{7}$$

and we can therefore conclude that only $g(y)$ can model strictly y-dependent features.

As desired, for vanishing h it still holds

$$p(y|x)\big|_{h(x,y)=0} \quad \propto \quad e^{g(y)} \propto p(y), \tag{8}$$

which implies independence, $y \perp x$. Thus, a non-trivial $h(x, y)$ models the $x \to y$ causal influence while a trivial $h(x, y) \equiv 0$ represents causal independence. We now want to make sure that

the more complicated x–dependence modeled by $h(x, y)$ is only introduced if it is strictly needed to explain the data. This way, we can clearly distinguish the causal scenarios $x \to y$ or $y \to x$ from the independent $x \perp y$ scenario. In fact, the distinction between g and h would be meaningless without a prior choice that favors independence between x and y. Hence, we assume g and h to be drawn from zero-centered Gaussian processes. In this way, without any information coming from the data, the most likely realizations of both of these functions are the identically-zero functions $g(y) = h(x, y) \equiv 0$. However, if the data exhibits strictly y-dependent features, meaning that the marginal distribution $p(y)$ is non-trivial, these features can only be represented by a non-zero $g(y)$.

Strictly y-dependent features are indeed clearly visible in the data. For example, we notice that in Fig 3 higher viral loads are by far more rare than lower viral loads. In this case, the data-inferred $g(y)$ is non-zero and shows this decreasing feature. Following the same reasoning, the most likely distribution for $h(x, y)$ in absence of data is identically-zero everywhere. Again, $h(x, y)$ will only be non-zero in case that the data triggers some coupling between x and y. Thus, the model favors independence of x and y (by favoring $h \equiv 0$) and the inferred density will be entangled in x and y only in the case the data exhibits this feature. This shows how the level of complexity of the model is adapted to the data automatically, without having to make any additional explicit model choice.

We again assume a Matérn-kernel-shaped correlation structure for the Gaussian process g, with covariance parameters $p_g := (a_g, k_g, \gamma_g)$. We set the priors on p_g as for p_f and learn these parameters from the data as well. Since the typical length scales and amplitudes of the variations of $h(x, y)$ in x and y directions are not in principle a priori similar, we assume the covariance for h to be shaped by a direct product of individual Matérn kernels in the x and y directions. For their corresponding prior parameters, $p_h = (p_h^{(x)}, p_h^{(y)})$ with $p_h^{(i)} = (a_h^{(i)}, k_h^{(i)}, \gamma_h^{(i)})$ and $i \in \{x, y\}$, we use similar hyper-priors as before, i.e. as for p_f and p_g, respectively. For more details on the prior choices we refer to the S1 Appendix. We call the ensemble of all these kernel parameters $p := (p_f, p_g, p_h)$. The details on how the Gaussian process for h with a Matérn kernel product covariance structure is set up is described in the S1 Appendix, where we describe a multi-dimensional density estimator that is agnostic to causal directions.

Lastly, we normalize the conditional PDF

$$p(y|x) = \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}} \left(\int \frac{e^{g(\tilde{y})+h(x,\tilde{y})}}{\int e^{h(\tilde{x},\tilde{y})}\mathrm{d}\tilde{x}} \mathrm{d}\tilde{y} \right)^{-1} \tag{9}$$

such that the full model density reads

$$\varrho(x, y) = \varrho_0 e^{f(x)} \frac{e^{g(y)+h(x,y)}}{\int e^{h(\tilde{x},y)}\mathrm{d}\tilde{x}} \left(\int \frac{e^{g(\tilde{y})+h(x,\tilde{y})}}{\int e^{h(\tilde{x},\tilde{y})}\mathrm{d}\tilde{x}} \mathrm{d}\tilde{y} \right)^{-1}. \tag{10}$$

A schematic representation of the forward causal model is shown in Fig 2. The assumed causal structure $x \to y$ is introduced in the model by the asymmetry between the roles of the x and y coordinates and the zero-centered Gaussian process priors on f, g, and h. Interchanging x and y leads to a model that follows the opposite causal direction $y \to x$. This allows to empirically distinguish these causal directions by calculating the model evidences for the two opposite scenarios, namely $x \to y$ and $y \to x$, as well as to test for $x \perp y$ by enforcing $h = 0$.

The evidence already inherenty penalizes more complicated models (see [36]), i.e. models that require a higher number of degrees of freedom. Using the ELBO as a criterion for model selection has therefore the additional benefit of regularizing the solutions. In other words, the evidence further helps the data analyst to univocally pick the lowest complexity model which

can best explain the data, in addition to the already self-adjusting level of complexity provided within the model itself.

Finally, we note that the choice of x and y is up to now completely arbitrary. Therefore, the method described in this work can be used to assess the causal direction of any continuous two-dimensional data distributions, given the data and some loose (prior) information about the correlation structure. the proposed method is thus very general and can be used for different datasets and analysis schemes with respect to those described in this paper. For example, it could be used to test the causal relation between the concentration of greenhouse gasses in the atmosphere and the average global temperature rise in Celsius degrees.

Likelihood

In order to construct the likelihood $\mathcal{P}(d|\varrho(\cdot,\cdot))$, we bin the data into a fine two dimensional grid over the x and y coordinates with 90×128 pixels, such that

$$n_{ij}(d) = \sum_{m=1}^{N} \int_{i\Delta x}^{(i+1)\Delta x} \mathrm{d}x \int_{j\Delta y}^{(j+1)\Delta y} \delta(x - x_m)\delta(y - y_m)\mathrm{d}y$$

contains the number of cases within the $(i, j)^{\text{th}}$ pixel of size $\Delta x = 1$ yr and $\Delta y \simeq 0.04 \log_{10}(\text{viral RNA copies}_i)/\text{ml}$. These counts n_{ij} are then compared with the model's expectations

$$\begin{aligned} \lambda_{ij} \quad &:= \lambda_{ij}(\varrho) = \int_{i\Delta x}^{(i+1)\Delta x} \mathrm{d}x \int_{j\Delta y}^{(j+1)\Delta y} \varrho(x, y)\mathrm{d}y \\ &\approx \Delta x \Delta y \;\; \varrho\left(\left(i + \frac{1}{2}\right)\Delta x, \left(j + \frac{1}{2}\right)\Delta y\right) \end{aligned} \tag{11}$$

via a Poisson likelihood

$$\mathcal{P}(d|\varrho) = \prod_{i,j} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} e^{-\lambda_{ij}}. \tag{12}$$

Inference

The full model involving the data d as well as all the unknown quantities, which compose the signal vector $s := (f, g, h, p)$, reads

$$\begin{aligned} \mathcal{P}(d, s) \quad &= \mathcal{P}(d|s)\,\mathcal{P}(s), \quad \text{where} \\ \mathcal{P}(d|s) \quad &= \mathcal{P}(d|\varrho[f, g, h]) \quad \text{and} \\ \mathcal{P}(s) \quad &= \mathcal{P}(f|p_f)\mathcal{P}(g|p_g)\mathcal{P}(h|p_h)\,\mathcal{P}(p_f)\,\mathcal{P}(p_g)\,\mathcal{P}(p_h). \end{aligned} \tag{13}$$

At this stage, we need to convert our causal model into an inference machine for the signal vector s. We do this conversion by reformulating the model in the language of information field theory [10, 11], transforming the coordinates of the signal vector $s = s(\xi)$ such that the prior on the new ξ coordinates becomes an uncorrelated Gaussian $\mathcal{P}(\xi) = \mathcal{G}(\xi, 1)$ as described by Knollmüller and Enßlin [37]. We then implement the resulting model using the Python package Numerical Information Field Theory (`NIFTy`) [38–40] and finally use `NIFTy`'s implementation of Metric Gaussian Variational Inference (MGVI) [41] to approximate the

posterior distribution in the new coordinates

$$\mathcal{P}(\xi|d) = \frac{\mathcal{P}(d|\xi)\mathcal{P}(\xi)}{\mathcal{P}(d)} \approx \mathcal{G}(\xi - \bar{\xi}_d, \Xi_d)$$

with a Gaussian which has posterior mean $\bar{\xi}_d$ and covariance Ξ_d, where the d suffix indicates the dataset used in the inference. This Gaussian posterior encodes the approximate result of the inference in the new coordinates. In order to translate this into the signal coordinates, we have to transform $\mathcal{P}(\xi|d)$ to $\mathcal{P}(s|d)$ using the relation $s = s(\xi)$. This relation is non-linear and the resulting PDF is neither Gaussian nor practical to obtain analytically. In order to evaluate moments from the posterior distributions of the desired quantities, MGVI provides ξ-samples drawn from the approximate Gaussian posterior $\xi \leftrightarrow \mathcal{G}(\xi - \bar{\xi}_d, \Xi_d)$. These ξ-samples can be converted via the coordinate transformation $s = s(\xi)$ into the signal space, where they represent (approximate) signal posterior samples. Making use of these posterior samples it is possible to calculate the posterior expectation values and model uncertainties of any desired quantity $q(s)$:

$$\begin{aligned} \bar{q} &:= \langle q(s) \rangle_{(s|d)} \approx \frac{1}{N_s} \sum_{i=1}^{N_s} q(s(\xi_i)) \\ \sigma_q^2 &= \langle (q(s) - \bar{q})^2 \rangle_{(s|d)}. \end{aligned} \tag{14}$$

Here, ξ_i denotes the ith of the N_s drawn samples. In particular, the posterior mean of the conditional PDF $p(y|x)$ and of any quantity which can be calculated from $p(y|x)$ can thereby be obtained, as well as the resulting uncertainties characterized via their uncertainty dispersion. In general, these posterior signal samples will not follow Gaussian statistics because the transformation is typically non-linear. Furthermore, since MGVI is a variational inference approach, the calculated uncertainties will be slightly smaller compared to the ones given from the accurate posterior. However given the complexity and size of the model, we need to use an approximate inference method as MGVI. For details about this methodology, as well as extensive performance and accuracy tests, we refer to Knollmüller and Enßlin [41].

Identifying causal directions with the Evidence Lower Bound

We now focus on understanding the causal relations given by the interplay between the variables. The proposed causal model should allow to discriminate between all possible causal structures, namely $x \to y$, $y \to x$, and $x \perp y$. Since we have a different model for each of these causal directions, we can use the Bayesian evidence to select the model that is better suited to the data. We do this once again by exploiting our variational inference scheme. It has been shown that the evidence is indeed a consistent and robust criterion for model selection [42, 43].

First, we need to define the independent model, i.e. a model for which age and viral load are regarded as statistically independent variables. Such a model is built by setting a very tight zero-centered prior on h, hence removing it from the model for all practical purposes. We can then estimate the Bayesian evidence both for the causal–hence dependent—models ($x \to y$ and $y \to x$) and compare it with the evidence of the independent model ($x \perp y$). More precisely, we calculate the so called Evidence Lower Bound (ELBO) as a proxy for an evidence. Using the posterior uncertainty covariance Ξ as well as the posterior samples provided by MGVI, we can compute the ELBO [36] for each model. This is lower than the exact logarithm of the evidence by the information difference (as measured in nits) between the exact and approximated posterior of a model. If too much information is not lost during MGVI, the ELBO should be a good

approximation of the exact log evidence. Furthermore, we can assume that deviations from the exact log evidence should be similar among different models, thereby reducing the effect of the MGVI approximations on differences between log-evidences. Thus, we can use the ELBO as a good proxy for the log model evidence ratios of similar models. The stochastic sampling steps performed in order to estimate the ELBO introduce a sampling uncertainty. This uncertainty can be in principle reduced by taking more samples, at the expense of larger computational costs. We state this numerical one-sigma uncertainty for all MGVI and ELBO based log-evidences. The $y \to x$ model is obtained by swapping the coordinates of the $x \to y$ model. The quantity of interest is then the logarithm of the evidence ratio of each causal model with the independent one,

$$\Delta E_{x \leftrightarrows y} = \log \frac{p(d|x \leftrightarrows y)}{p(d|x \perp y)}, \tag{15}$$

where $x \leftrightarrows y$ denotes either $x \to y$ or $y \to x$. $\Delta E_{x \to y}$ indicates the log evidence in favor of the causal model $x \to y$ with respect to the independent one and similarly $\Delta E_{y \to x}$ the one for $y \to x$. Comparing $\Delta E_{x \to y}$ with $\Delta E_{y \to x}$ also allows to discriminate between the two possible causal directions in the dataset. We identify the preferred causal direction underlying the data in subsec. Causal directions and bias, where we also discuss its implications.

Data

We make use of RT-PCR viral load data collected from the Charité Institute of Virology and Labor in Berlin, Fig. 6 in Jones et al. [9]. The data was acquired with two different PCR instruments, Roche cobas 6800/8800 (cobas dataset, which we denote by d_C and is comprised of ≈2200 data points) and Roche LightCycler 480 II (LC 480 dataset, which we denote by d_L comprised of ≈1350 data points). In the following, we will show the difference between the two datasets. As can be seen from the count difference in the raw data plots as of Fig. 6 of [9] and in the histogramed data in Figs 3 and 4, for low viral loads ($y \lesssim 5$ in units of $\log_{10}$(RNA copies/ml)), the LC 480 dataset shows a roughly uniform count distribution in the whole viral load domain. In contrast, the cobas dataset exhibits an increasing number of counts in the $y \in [2.0, 3.8]$ viral load domain followed by a descending trend in counts in the $y \in [3.8, 5.0]$ region. For this reason and in order to better understand the possible shortcomings of both instruments, we analyzed the data in two different ways.

Since the major differences between the datasets arise for viral load values $y \in [3.8, 5.0]$ we define two lower thresholds for the viral load ($y_{\min} := 3.8$ and $y'_{\min} := 5.4 \simeq \log_{10}(250000)$ in units of $\log_{10}$(viral RNA copies/ml)) and discard any data point for which the viral load is lower than $y_{\min}$ and $y'_{\min}$, respectively. We set the lower threshold $y_{\min}$ at the value for which the number of counts of the cobas dataset is maximum (see Fig 1). Below this threshold, the counts' density lowers dramatically. As for $y'_{\min}$, the value of 250000 in units of viral RNA copies/ml indicates the threshold for the isolation of infectious virus in cell cultures at more than 5% probability as described by Wörfel et al. [44]. We then analyze the cobas and LC 480 datasets first neglecting the data below $y_{\min}$ threshold and then below $y'_{\min}$. For the sake of simplicity, we will denote the cobas and LC 480 datasets with $y_{\min}$ and $y'_{\min}$ as a lower threshold as d_C, d'_C, d_L, and d'_L, respectively. This way we can highlight the differences between the two datasets and investigate possible sources of systematic errors in the viral load measurement instruments or different selection effects in the data collection process.

The datasets are not explicitly provided by the authors. Therefore, we acquired the data by means of a plot digitizer algorithm from Fig. 6 of [9]. Since the age coordinates are not labeled precisely, but only a rough interval $\Delta_{\text{age}} \sim 0 - 100$ yr is provided, we do not expect the acquired

data points to be accurate—especially in the age domain. For this reason the obtained age axis could be affected by a global shift up to 5 − 10 years in any direction. But, since our model is translation invariant, this kind of systematic bias in the data extraction process (an overall shift in the age values for all data points) does not affect the causal inference machinery. Nonetheless, the reader should keep this in mind when interpreting the results concerning the conditional PDF of the viral load given the age. Concerning the viral load coordinates, for which more precise units were given, the data should be regarded as more reliable. For our aim of building a method to analyze age and viral load data in a non-parametric and causal fashion providing uncertainty estimates, this level of accuracy is sufficient. The results we present from here on are given for the measured coordinate values without considering any uncertainty with respect to the real quantities (age, viral load). Nevertheless, we do not believe systematic or random error contributions in the data extraction procedure to significantly affect our results since the shapes of the learned distributions are translation invariant.

Results

In the following, we discuss the main results and their relevance for the infectivity of different age groups.

Age dependence of the viral load

For both lower thresholds and datasets d_C, d_L, we estimate the model parameters s by means of the MGVI algorithm implemented in `NIFTy`. In Figs 3 and 4 we show the data (left panel) and the correspondent reconstructed underlying densities $\varrho(x, y)$ (central panel).

A multi-modal age distribution is clearly visible as well as an overall decrease for growing viral loads in all of the reconstructed densities. The conditional PDF $p(y|x)$ of the viral load y given the age x is shown in Fig 5. In the conditional probability, the multi-modal age structure displayed by the density is not visible since it has been absorbed by $\varrho(x)$. What this effectively means is that age-only selection effects—e.g., testing one or many specific age groups more than others or demographics in general—have been modeled, and the resulting conditional probability distribution does not depend on such effects.

For all datasets and ages, a general descending trend in the viral load probability distribution is clearly visible. The reconstruction based on the d_C dataset exhibits significant differences in the viral load for different ages (Fig 5, first panel). For infected patients approximately above the age of 60, the distribution exhibits a distinct maximum for viral loads of $y \simeq 8$ (in units of $\log_{10}$(RNA copies/ml)). Furthermore, we show that this feature is indeed triggered by the data, and is not just the result of an over-fit of sample noise. To do so, we apply a random permutation r to the viral load values in the d_C dataset, the only one that exhibits a possible $x \to y$ causal structure. We then analyze the randomized dataset $d_C^r = \{(i, x_i, y_{r(i)})\}_{i=1}^N$ in the same way as seen for $d_C = \{(i, x_i, y_i)\}_{i=1}^N$. The resulting conditional PDF $p^r(y|x)$ reconstructed from the randomized dataset (Fig 5) does not exhibit any clear age-dependent structure in the viral load, indicating that the differences seen in the real data are not just a shot noise effect.

Causal directions and bias

As discussed in Subsec. Causal structure, we can use ELBO ratios to identify causal relations. In our analysis, this corresponds to choosing between the $x \to y$, $y \to x$, and $x \perp y$ models. Identifying the causal direction underlying the data will allow to show that selection effects distorting the expected causal relation $x \to y$ are subdominant and that for the cobas dataset d_C we indeed see evidence for an age dependence of the viral load distribution, $p(y|x) \neq p(y)$.

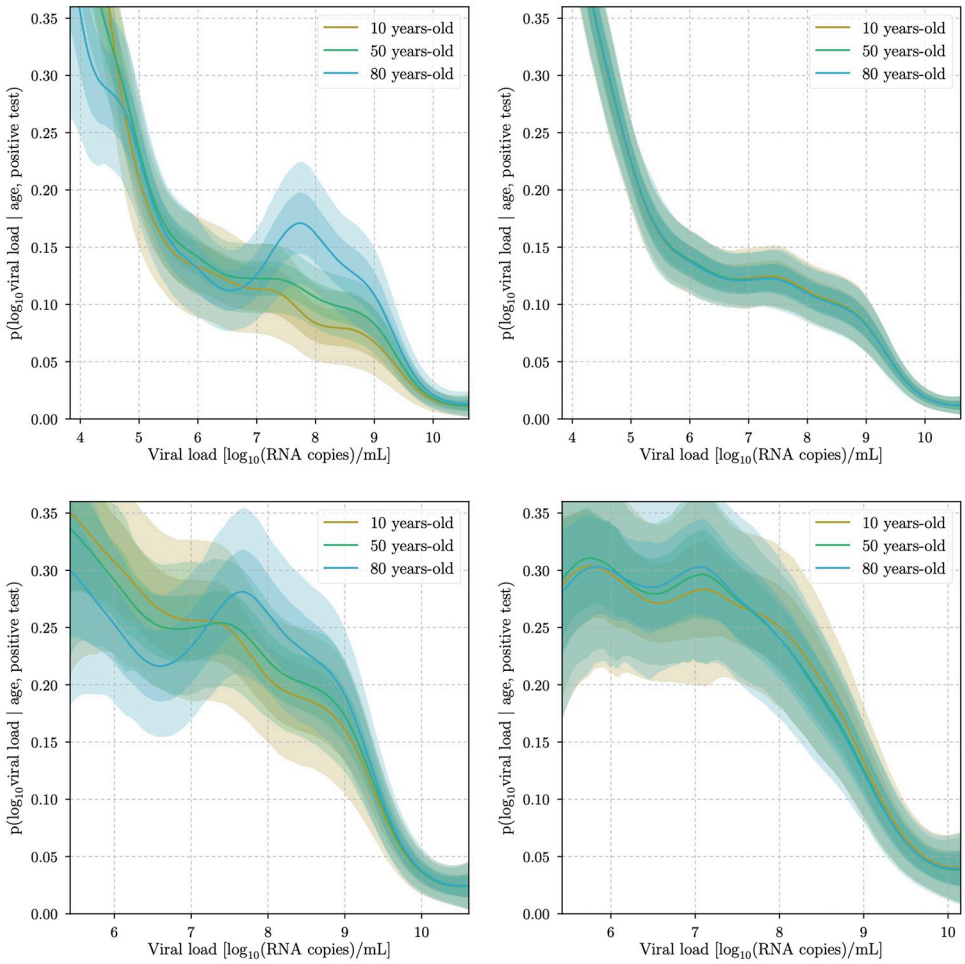

Fig 5. Viral-load probabilities for specific ages. Viral load conditional PDF $p(y|x)$ for specific ages $x \in \{10, 50, 80\}$. Panel one and two (top) display the results for the cobas dataset d_C and for the randomized cobas dataset d_C respectively. The latter serves as a null test, since the randomization erases any (causal) relation between the age (x) and the viral load (y) other than shot noise. Panel three and four (bottom) show $p(y|x)$ for the cobas dataset d'_C and for the LC 480 dataset d'_L respectively, both with the higher viral load threshold y'_{min}. The shaded regions represent 1σ and 2σ uncertainty contours of the approximate posterior.

https://doi.org/10.1371/journal.pone.0275011.g005

As a matter of fact, the log-evidence ratio for the d_C dataset is $\Delta E_{C,x\to y} = 4.6 \pm 1.0$ for d_C, which clearly favors the dependent model, but this value decreases to $\Delta E'_{C,x\to y} = -1.5 \pm 1.0$ when considering d'_C, hence y'_{min} as a lower threshold. We highlight that a log-evidence difference between the two compared models of 1 unit corresponds to a factor of $e \approx 2.7$ for the Bayesian odds ratio between the two. Thus, $\Delta E_{C,x\to y} = 4.6 \pm 1.0$ implies given equal model priors, $p(x \to y) = p(x \perp y)$, a posterior model odds ratio of $p(x \to y|d) : p(x \perp y|d) = e^{3.5\pm0.7} \approx 99.5^{[270.4]}_{[36.6]}$ in favor of a causal dependence between viral load and age for the cobas dataset with the lower threshold y_{min}.

For the opposite causal direction we get $\Delta E_{C,y\to x} = -0.2 \pm 1.1$, which shows that there is no strong $y \to x$ structure in the data. For the LC dataset d_L these evidence differences with respect to the independent model become $\Delta E_{L,x\to y} = -4.6 \pm 1.0$ and $\Delta E'_{L,x\to y} = -3.4 \pm 1.0$ respectively for the two thresholds. Therefore the independent model is favored in both cases. This shows that the cobas (d'_C) and LC 480 (d'_L) datasets are in agreement for viral loads which are higher

than $y'_{\min}$, but in case we include the data lying in the viral load region $y \in [y_{\min}, y'_{\min}]$, the causal age-dependent structure becomes visible in d_C and is not negligible anymore.

It is known that model evidences can vary strongly with different data realizations. Moreover, in order to calculate the evidences, we invoked approximations and stochastic calculation steps. Thus, proper null-tests are required in order to validate and calibrate the evidence ratio calculation. We provide such null-tests by repeating the data randomization step described above several times, thereby producing many randomized datasets. By construction, these dataset should not exhibit any causal structure. Indeed, for the 10 randomized-dataset realizations performed, we find much lower log-evidence ratios between dependent (causal) and independent models with respect to the ones found for the original cobas dataset d_C. The average difference between these tests is $\langle \Delta E_{\text{random},x\to y} \rangle_{\text{randomizations}} = -6.0 \pm 1.0$. Since none of the randomized dataset realizations reaches comparably high log-evidence ratios with respect to the independent model, all these findings support robustly the argument that the dependent model is a more suited description of the d_C dataset. As previously discussed, this result also supports the assumption that no additional variable z is confounding $x \leftarrow z \rightarrow y$ the bivariate causal direction $x \rightleftarrows y$. Had there been a strong confounder z, we would have expected a more complicated data distribution that could have possibly been detected when comparing the actual data with the randomized datasets. Even if this did not appear to be the case, we cannot completely rule out the presence of a weak confounder.

The results of the evidence calculations are displayed in Fig 6. This figure also indicates that the independent model interpretation of the data is favored for all other datasets and threshold combinations (except for d_C), since the evidence for the independent model is always higher.

These contradicting results for the two datasets (or thresholds) d_C and d'_C might have several possible explanations. First, they could be caused by a potential accuracy loss of the PCR devices below certain viral load values, as suggested for the cobas dataset d_C below $y'_{\min}$ in [9], hence for the Roche cobas 6600/8800 PCR system. It could also mean that the opposite is happening and the Roche LC 480 PCR device is less sensitive than the Roche cobas 6600/8800 in the $y \in [3.8, 5.0]$ region. This possibility would be supported by the fact that the cobas dataset exhibits a clear age dependent structure in such viral load region, but the (swab) data processed

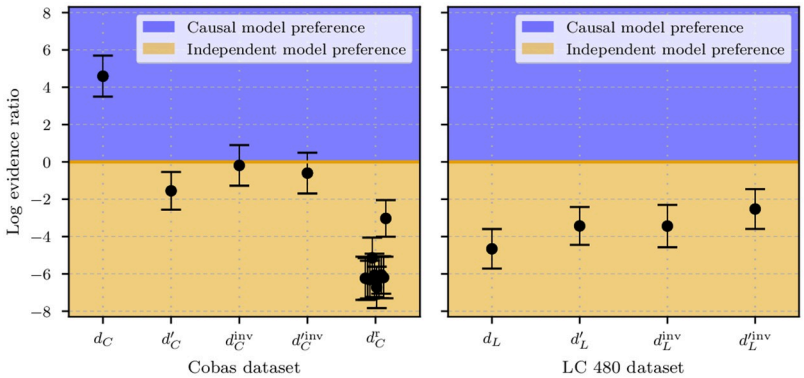

Fig 6. Evidence results. Natural logarithm of the evidence ratio with respect to the correspondent independent model for the $x \to y$ cobas datasets d_C and d'_C, for the $y \to x$ causal model d_C^{inv} and d'^{inv}_C (left panel). The log-evidence ratios labeled with d^r_C display the (y-)randomized datasets (for 10 different realizations). The same is also shown for the LC 480 dataset (right). The error bars represent the numerical uncertainty associated to the stochastic estimate of the ELBO. A positive logarithm of the evidence ratio denotes a preference for the given causal model with respect to the correspondent independent model and vice versa. The strength of this preference can be determined by taking the exponential of the log-evidence ratio between the causal model and the correspondent independent one. For example, a log-evidence ratio of 5 corresponds to a $e^5 \approx 141$-fold preference for the examined causal model.

https://doi.org/10.1371/journal.pone.0275011.g006

by the cobas PCR device contains no information on the patients' ages. Hence it would be surprising that a systematic effect in the measurements could introduce an age dependence on the viral load distribution. Furthermore, this pattern–that older patients exhibit higher viral loads—is plausible from a medical perspective.

Nevertheless, we cannot exclude the possibility that selection effects have been introduced in the data. We have already shown that "viral load causing age" effects ($y \to x$) are subdominant. Nevertheless, selection effects could still have been introduced for instance by collecting age and viral load subsamples from a "viral-load biased" population sample. This could happen, e.g., if symptomatic patients had been predominantly tested. Since children are less likely to show symptoms than adults, the sample would then include mainly those children with higher viral loads.

And finally, a combination of such competing effects could have affected the results and thereby imprinted a spurious age dependence to the viral-load distribution. Given only the statistical data in our possession, this possibility cannot be ruled out completely.

Impact on infectivity

After having established a potential age difference in the viral load distribution for d_C, we investigate whether this difference—if real—would be relevant for the infection dynamics. For this purpose, we need to link the viral load to the infectivity $I(y)$ of the virus, i.e. the probability of transmitting the infection. Infectivity can be measured in different ways. In our analysis, we choose the projected virus isolation success based on probit distribution described in [44] as a proxy for infectivity. This represents the infection success rate for cell cultures exposed to saliva with different viral load y and can be seen as the blue curve in Fig 7, labeled with "Original". We will from now on refer to this parameter as "infectivity proxy" and indicate it with $I(y)$.

The projected infectivity as a function of the age is then given by the expectation value of the infectivity parameter over the conditional PDF $p(y|x)$

$$I(x) \coloneqq \langle I(y) \rangle_{p(y|x)} = \int I(\tilde{y}) p(\tilde{y}|x) \mathrm{d}\tilde{y}. \tag{16}$$

The result, together with the uncertainties resulting from our PDF modeling are shown in Fig 7. No relative differences larger than **0.18** for the (projected) infectivity of the different age groups is found, with typical values of $I(x) \approx 0.3$ for all datasets. This means that at most a 50% difference in infectivity due to different viral load between different age groups—but more likely a smaller one—should be expected.

In order to characterize the uncertainty resulting from our $I(y)$ modeling, due to the uncertainty of the original determination of this function and due to the uncertainty in the identification of viral loads with different instruments, we repeat the analysis while shifting the original $I(y)$ curve by one order of magnitude upwards and downwards in y. The resulting maximal relative difference in the infectivity of the different age groups is $\simeq 0.3$. Thus, even though our model allows us to show that infectivity exhibits an age dependence if the $\boldsymbol{d}_C$ dataset with $y_{\min}$ provides a valid picture, the viral load differences between different age groups though are not strong enough to impact on the infection dynamics at a level that justifies regarding any age group as noninfectious or even significantly less infectious.

Conclusions

In order to investigate the controversial results reported in the literature, we developed a causal model to assess the dependence of viral loads of patients infected with COVID-19 on age. The

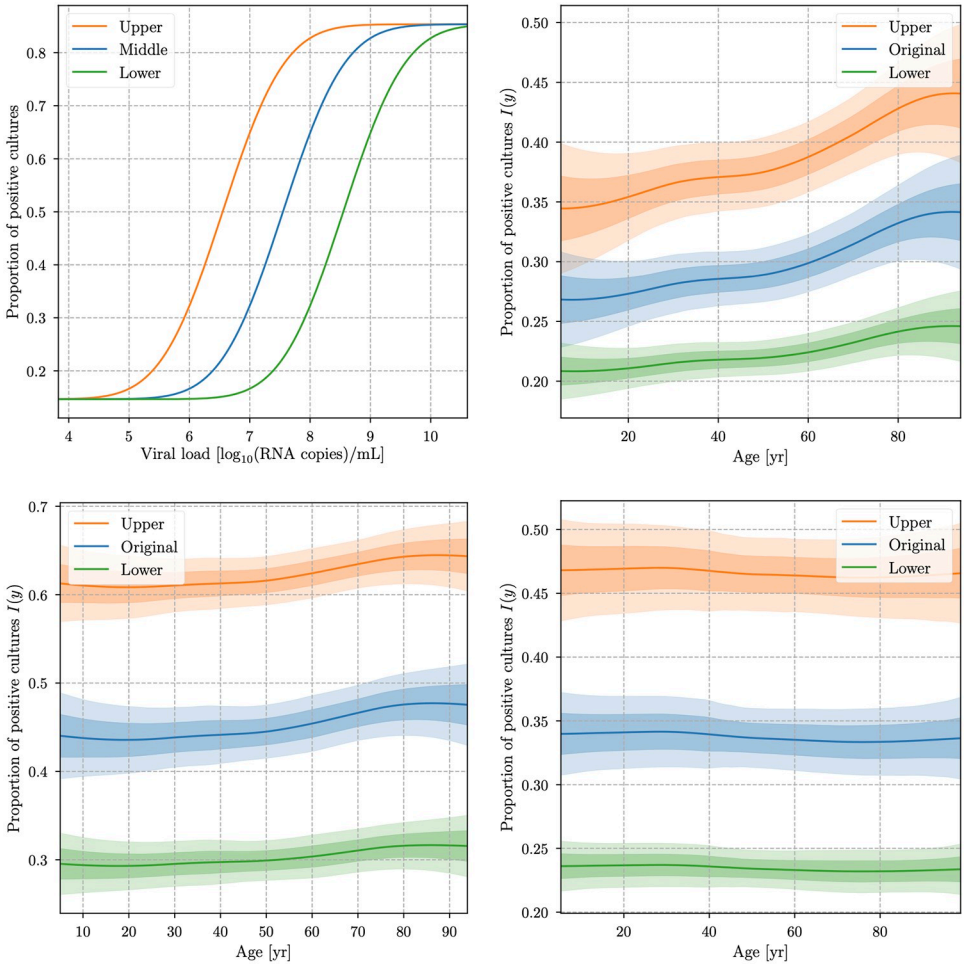

Fig 7. Infectivity proxy vs. age. $I(y)$ obtained from Fig. 1g in [44] and fit with a probit function (top left, Original), then projected for different patients' ages for the cobas dataset d_C (top right), d'_C (bottom left) and for the LC 480 dataset d_L (bottom right). The upper and lower curves are obtained by translating $I(y)$ of one unit in viral load as in $I(y-1)$ and $I(y+1)$ in order to give upper and lower bounds similar as those shown in Fig. 1g in [44]. The shaded regions show 1σ and 2σ uncertainty contours of the approximate posterior.

https://doi.org/10.1371/journal.pone.0275011.g007

developed model is capable of reconstructing two-dimensional density distributions from data counts and to learn causal directions. The model complexity is set in a user-independent fashion making the results more robust and consistent. Furthermore, its causal nature allows to make predictions about non-directly measured quantities (e.g. the infectivity assessment in Sec. Results and Fig 7) and to additionally test for bias in the data-collection process.

Although the benefits of a causal analysis were already discussed in Pearl's original work [32] and are usually well recognized, causal inference is not so commonly applied to real-world data science, often because it requires the implementation of complicated ad-hoc models. With the hope of making it useful for a wide range of data-science applications we make our causal model freely available. This model is very flexible and generic (as it only requires a set of x and y data pairs and mild priors on their correlation structures) and it can therefore be used in future epidemiological studies as well as in completely different fields. We provide the source code under an open source license for usage in further studies and applications at https://gitlab.mpcdf.mpg.de/ift/public/causal_age_viral_load_model. As a side product, we

also developed a causal-direction-agnostic density estimator, which is described in more detail in the S1 Appendix.

Using our novel method to model causal relations non-parametrically, we have reanalyzed the SARS-CoV-2 age and viral load data presented in [9]. In doing so, we have found statistically significant differences in the viral load distribution of different age groups when regarding the cobas dataset d_C for viral loads within the interval of $10^{3.8}$ to $10^{5.0}$ in units of viral RNA copies/ml of sample or entire swab specimen as reliable. These differences become irrelevant if this region is ignored in the analysis.

We cannot completely exclude that selection effects in the data-collection process may have introduced an apparent causal relation between viral load and age, but the observed trend—a statistically-significant increase in the viral load with age—fits with the generally accepted notion that the immune system response gets weaker with age. Assuming this trend to be real, we showed, however, that its expected impact on the infectivity of different age groups is at most moderate. For this reason we cannot exclude any age group from being considered as a potentially significant source of infection.

The region of the cobas dataset relevant for this trend is described in [9] as containing an artifact, suggesting that the correct interpretation of the data is that viral load, hence infectivity, is predominantly age independent. Here, we want to point out that other studies on the age dependence of the viral load present in the literature [15, 18] make the opposite claim. Moreover, in their most recent publication, Jones et al. [31] acknowledge an age dependence of the viral load. This dependence is quantitatively similar to the one we have detected with our method. Furthermore, the causal evidence tests presented in Sec. Results favor considering the age dependence of the viral load as a real effect and not just as an artifact. These tests also disfavor the reverse-causal-direction model (here: that the viral load of a patient "causes" its age), which would indicate that strong selection effects have affected the data-collection process. We introduced these tests as a new tool to detect potential systematic effects in similar datasets.

In conclusion, the results of our analysis ultimately confirm some of the findings in the literature—i.e. that the viral load is only modestly dependent on the age—but with a much higher sensitivity and robustness. Of central importance are the methods here developed. While being tailored to describe the Covid-19 pandemic data, they can be easily adapted for more general purposes and can prove very useful also for future pandemics or for new—and possibly more infective—mutations of SARS-CoV-2.

Supporting information

S1 Appendix. Matérn-kernel density reconstruction.
(PDF)

Acknowledgments

We thank Dr. Matthia Sabatelli (Montefiore Institute, University of Liège) for feedback on preliminary versions of this work and the Information Field Theory group at the Max Planck Institute for Astrophysics for the fruitful discussions.

Author Contributions

Conceptualization: Matteo Guardiani, Philipp Frank, Andrija Kostić, Torsten Enßlin.

Data curation: Matteo Guardiani.

Formal analysis: Matteo Guardiani, Philipp Frank, Andrija Kostić, Torsten Enßlin.

Investigation: Matteo Guardiani.

Methodology: Matteo Guardiani, Gordian Edenhofer, Jakob Roth, Torsten Enßlin.

Project administration: Matteo Guardiani.

Software: Matteo Guardiani, Andrija Kostić, Gordian Edenhofer, Jakob Roth.

Supervision: Philipp Frank, Torsten Enßlin.

Writing – original draft: Matteo Guardiani, Berit Uhlmann, Torsten Enßlin.

Writing – review & editing: Matteo Guardiani, Philipp Frank, Andrija Kostić, Gordian Edenhofer, Jakob Roth, Berit Uhlmann, Torsten Enßlin.

References

1. Colson P, Tissot-Dupont H, Morand A, Boschi C, Ninove L, Esteves-Vieira V, et al. Children account for a small proportion of diagnoses of SARS-CoV-2 infection and do not exhibit greater viral loads than adults. European Journal of Clinical Microbiology & Infectious Diseases. 2020; 39(10):1983–1987. https://doi.org/10.1007/s10096-020-03900-0
2. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q, et al. Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study. Bmj. 2020; 369. https://doi.org/10.1136/bmj.m1443 PMID: 32317267
3. Pujadas E, Chaudhry F, McBride R, Richter F, Zhao S, Wajnberg A, et al. SARS-CoV-2 viral load predicts COVID-19 mortality. The Lancet Respiratory Medicine. 2020; 8(9):831–934. https://doi.org/10.1016/S2213-2600(20)30354-4 PMID: 32771081
4. Fajnzylber J, Regan J, Coxen K, Corry H, Wong C, Rosenthal A, et al. SARS-CoV-2 viral load is associated with increased disease severity and mortality. Nature Communications. 2020; 11(1):5493. https://doi.org/10.1038/s41467-020-19057-5 PMID: 33127906
5. Maltezou HC, Raftopoulos V, Vorou R, Papadima K, Mellou K, Spanakis N, et al. Association Between Upper Respiratory Tract Viral Load, Comorbidities, Disease Severity, and Outcome of Patients With SARS-CoV-2 Infection. The Journal of Infectious Diseases. 2021; https://doi.org/10.1093/infdis/jiaa804 PMID: 33388780
6. Yonker LM, Neilan AM, Bartsch Y, Patel AB, Regan J, Arya P, et al. Pediatric severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): clinical presentation, infectivity, and immune responses. The Journal of pediatrics. 2020; 227:45–52. https://doi.org/10.1016/j.jpeds.2020.08.037 PMID: 32827525
7. Jefferson T, Spencer E, Brassey J, Heneghan C. Viral cultures for COVID-19 infectivity assessment. Systematic review. medRxiv. 2020;.
8. Kawasuji H, Takegoshi Y, Kaneda M, Ueno A, Miyajima Y, Kawago K, et al. Transmissibility of COVID-19 depends on the viral load around onset in adult and symptomatic patients. PLoS ONE. 2020; 15(12). https://doi.org/10.1371/journal.pone.0243597 PMID: 33296437
9. Jones TC, Mühlemann B, Veith T, Biele G, Zuchowski M, Hoffmann J, et al. An analysis of SARS-CoV-2 viral load by patient age. https://virologie-ccm.charite.de/fileadmin/user_upload/microsites/m_cc05/virologie-ccm/dateien_upload/Weitere_Dateien/Charite_SARS-CoV-2_viral_load_2020-06-02.pdf
10. Enßlin TA, Frommert M, Kitaura FS. Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis. Physical Review D. 2009; 80(10):105005. https://doi.org/10.1103/PhysRevD.80.105005
11. Enßlin TA. Information theory for fields. Annalen der Physik. 2018; p. 1800127.
12. Leike RH, Glatzle M, Enßlin TA. Resolving nearby dust clouds. Astronomy & Astrophysics. 2020; 639: A138. https://doi.org/10.1051/0004-6361/202038169
13. Arras P, Frank P, Haim P, Knollmüller J, Leike R, Reinecke M, et al. The variable shadow of M87. arXiv preprint arXiv:200205218. 2020;.
14. Kurthen M, Enßlin T. A Bayesian Model for Bivariate Causal Inference. Entropy. 2020; 22(1):46. https://doi.org/10.3390/e22010046
15. Mahallawi WH, Alsamiri AD, Dabbour AF, Alsaeedi H, Al-Zalabani AH. Association of Viral Load in SARS-CoV-2 Patients With Age and Gender. Frontiers in Medicine. 2021; 8:39. https://doi.org/10.3389/fmed.2021.608215 PMID: 33585523

16. Jacot D, Greub G, Jaton K, Opota O. Viral load of SARS-CoV-2 across patients and compared to other respiratory viruses. Microbes and infection. 2020; 22(10):617–621. https://doi.org/10.1016/j.micinf.2020.08.004 PMID: 32911086

17. Kleiboeker S, Cowden S, Grantham J, Nutt J, Tyler A, Berg A, et al. SARS-CoV-2 viral load assessment in respiratory samples. Journal of Clinical Virology. 2020; 129:104439. https://doi.org/10.1016/j.jcv.2020.104439 PMID: 32674034

18. Euser S, Aronson S, Manders I, van Lelyveld S, Herpers B, Jansen R, et al. SARS-CoV-2 viral load distribution in different patient populations and age groups reveals that viral loads increase with age. medRxiv. 2021; https://doi.org/10.1093/ije/dyab145 PMID: 34999848

19. Costa R, Bueno F, Albert E, Torres I, Carbonell-Sahuquillo S, Barres-Fernandez A, et al. Upper respiratory tract SARS-CoV-2 RNA loads in symptomatic and asymptomatic children and adults. medRxiv. 2021;.

20. Heald-Sargent T, Muller WJ, Zheng X, Rippe J, Patel AB, Kociolek LK. Age-related differences in nasopharyngeal severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) levels in patients with mild to moderate coronavirus disease 2019 (COVID-19). JAMA pediatrics. 2020; 174(9):902–903. https://doi.org/10.1001/jamapediatrics.2020.3651 PMID: 32745201

21. Chung E., Chow E., Wilcox N., Burstein R., Brandstetter E., Han P., et al. Comparison of symptoms and RNA levels in children and adults with SARS-CoV-2 infection in the community setting. *JAMA Pediatrics*. 175, e212025–e212025 (2021) https://doi.org/10.1001/jamapediatrics.2021.2025 PMID: 34115094

22. Baggio S., Arnaud G., Yerly S., Bellon M., Wagner N., Rohr M., et al. SARS-CoV-2 Viral Load in the Upper Respiratory Tract of Children and Adults With Early Acute COVID-19. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America, 73(1), 148–150.

23. Polese-Bonatto M., Sartor I., Varela F., Giannini G., Azevedo T., Kern L., et al. Children Have Similar Reverse Transcription Polymerase Chain Reaction Cycle Threshold for Severe Acute Respiratory Syndrome Coronavirus 2 in Comparison With Adults. *The Pediatric Infectious Disease Journal*. 40, e413 (2021) https://doi.org/10.1097/INF.0000000000003300 PMID: 34596626

24. Ade C., Pum J., Abele I., Raggub L., Bockmühl D. & Zöllner B. Analysis of cycle threshold values in SARS-CoV-2-PCR in a long-term study. *Journal Of Clinical Virology*. 138 pp. 104791 (2021) https://doi.org/10.1016/j.jcv.2021.104791 PMID: 33725648

25. Cendejas-Bueno E., Romero-Gómez M., Escosa-García L., Jiménez-Rodríguez S., Mingorance J., García-Rodríguez J., Group S. & Others Lower nasopharyngeal viral loads in pediatric population. The missing piece to understand SARS-CoV-2 infection in children?. *Journal Of Infection*. 83, e18–e19 (2021) https://doi.org/10.1016/j.jinf.2021.06.009 PMID: 34133963

26. Bullard J., Funk D., Dust K., Garnett L., Tran K., Bello A., et al. Infectivity of severe acute respiratory syndrome coronavirus 2 in children compared with adults. *Cmaj*. 193, E601–E606 (2021) https://doi.org/10.1503/cmaj.210263 PMID: 33837039

27. Bellon M., Baggio S., Jacquerioz Bausch F., Spechbach H., Salamun J., Genecand C., et al. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral load kinetics in symptomatic children, adolescents, and adults. *Clinical Infectious Diseases*. 73, e1384–e1386 (2021) https://doi.org/10.1093/cid/ciab396 PMID: 33949655

28. Maltezou HC, Magaziotou I, Dedoukou X, Eleftheriou E, Raftopoulos V, Michos A, et al. Children and adolescents with SARS-CoV-2 infection: epidemiology, clinical course and viral loads. The Pediatric Infectious Disease Journal. 2020; 39(12):e388–e392. https://doi.org/10.1097/INF.0000000000002899 PMID: 33031141

29. L'Huillier AG, Torriani G, Pigny F, Kaiser L, Eckerle I. Culture-competent SARS-CoV-2 in nasopharynx of symptomatic neonates, children, and adolescents. Emerging infectious diseases. 2020; 26(10):2494. https://doi.org/10.3201/eid2610.202403 PMID: 32603290

30. Zachariah P, Halabi KC, Johnson CL, Whitter S, Sepulveda J, Green DA. Symptomatic infants have higher nasopharyngeal SARS-CoV-2 viral loads but less severe disease than older children. Clinical Infectious Diseases. 2020; 71(16):2305–2306. https://doi.org/10.1093/cid/ciaa608 PMID: 32433729

31. Jones TC, Biele G, Mühlemann B, Veith T, Schneider J, Beheim-Schwarzbach J, et al. Estimating infectiousness throughout SARS-CoV-2 infection course. Science. 2021; https://doi.org/10.1126/science.abi5273 PMID: 34035154

32. Pearl J. Causality: Models, Reasoning and Inference. Cambridge University Press, 2009.

33. Spirtes P., Glymour C. & Scheines R. Causation, Prediction, and Search, 2nd Edition. The MIT Press, 2001, 12

34. Mooij J., Peters J., Janzing D., Zscheischler J. & Schölkopf B. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal Of Machine Learning Research*. 17, 1–102 (2016), http://jmlr.org/papers/v17/14-518.html

35. Li L., White L.F. Bayesian back-calculation and nowcasting for line list data during the COVID-19 pandemic. *PLoS Computational Biology* 17, 1–22 (2021,7), https://doi.org/10.1371/journal.pcbi.1009210 PMID: 34252078

36. Kostić A, Leike R, Hutschenreuter S, Kurthen M, Enßlin TA. Bayesian Causal Inference with Information Field Theory. Manuscript in preparation (2022).

37. Knollmüller J, Enßlin TA. Encoding prior knowledge in the structure of the likelihood. arXiv e-prints. 2018; p. arXiv:1812.04403.

38. Selig M, Bell MR, Junklewitz H, Oppermann N, Reinecke M, Greiner M, et al. NIFTy—Numerical Information Field Theory. A versatile Python library for signal inference. Astron. & Astrophys. 2013; 554: A26. https://doi.org/10.1051/0004-6361/201321236

39. Steininger T, Dixit J, Frank P, Greiner M, Hutschenreuter S, Knollmüller J, et al. NIFTy 3—Numerical Information Field Theory: A Python Framework for Multicomponent Signal Inference on HPC Clusters. Annalen der Physik. 2019; 531(3):1800290. https://doi.org/10.1002/andp.201800290

40. Arras P, Baltac M, Ensslin TA, Frank P, Hutschenreuter S, Knollmueller J, et al.. NIFTy5: Numerical Information Field Theory v5; 2019.

41. Knollmüller J, Enßlin TA. Metric Gaussian Variational Inference. arXiv e-prints. 2019; p. arXiv:1901.11033.

42. Cherief-Abdellatif B. Consistency of ELBO maximization for model selection. *Proceedings Of The 1st Symposium On Advances In Approximate Bayesian Inference*. 96 pp. 11–31 (2019,12,2).

43. Knuth K., Habeck M., Malakar N., Mubeen A. & Placek B. Bayesian evidence and model selection. *Digital Signal Processing*. 47 pp. 50–67 (2015), Special Issue in Honour of William J. (Bill) Fitzgerald https://doi.org/10.1016/j.dsp.2015.06.012

44. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020; 581(7809):465–469. https://doi.org/10.1038/s41586-020-2196-x PMID: 32235945