Gender and sex bias in COVID-19 epidemiological data through the lens of causality

Natalia Díaz-Rodríguez, Rūta Binkytė, Wafae Bakkali, Sannidhi Bookseller, Paola Tubaro, Andrius Bacevičius, Sami Zhioua, Raja Chatila

Information Processing and Management 60 (2023) 103276

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Gender and sex bias in COVID-19 epidemiological data through the lens of causality

Natalia Díaz-Rodríguez [a,1,*], Rūta Binkytė [b,1], Wafae Bakkali [c,2], Sannidhi Bookseller [d], Paola Tubaro [e], Andrius Bacevičius [f], Sami Zhioua [b], Raja Chatila [g]

[a] DaSCI Andalusian Institute in Data Science and Computational Intelligence, CITIC, Dpt. of Computer Science and Artificial Intelligence, University of Granada, Spain
[b] INRIA, École Polytechnique, IPP, Paris, France
[c] Amazon Machine Learning Solutions Lab, Amazon Web Services, Paris, France
[d] EPITA College, Le Kremlin-Bicêtre, France
[e] LISN-TAU, CNRS, University Paris-Saclay, Inria, France
[f] OSE Immunotherapeutics, Paris, France
[g] ISIR (Institute of Intelligent Systems and Robotics), Sorbonne University, Paris, France

ARTICLE INFO

Keywords:
Explainability
Causality
Causal fairness
COVID-19
Sex
Gender
Equality
Artificial intelligence
Healthcare

ABSTRACT

The COVID-19 pandemic has spurred a large amount of experimental and observational studies reporting clear correlation between the risk of developing severe COVID-19 (or dying from it) and whether the individual is male or female. This paper is an attempt to explain the supposed male vulnerability to COVID-19 using a causal approach. We proceed by identifying a set of confounding and mediating factors, based on the review of epidemiological literature and analysis of sex-dis-aggregated data. Those factors are then taken into consideration to produce explainable and fair prediction and decision models from observational data. The paper outlines how non-causal models can motivate discriminatory policies such as biased allocation of the limited resources in intensive care units (ICUs). The objective is to anticipate and avoid disparate impact and discrimination, by considering causal knowledge and causal-based techniques to compliment the collection and analysis of observational big-data. The hope is to contribute to more careful use of health related information access systems for developing fair and robust predictive models.

1. Introduction

Sex and gender disparity was noticed in many cases of Coronavirus disease 2019 (COVID-19). In this article we follow the definition proposed by Ahmed and Dumanski (2020) distinguishing between *sex* as a set of biological attributes, and *gender* as a social–psychological category. As it is later demonstrated, both might have an impact on COVID-19 mortality rates. We also note that in this study we consider binary values for gender, although an in depth analysis of gender roles would potentially yield a more complex picture. The disease is reported to be deadlier for infected men than women with a 2.8% fatality rate in Chinese men versus 1.7% in women (Gebhard, Regitz-Zagrosek, Neuhauser, Morgan, & Klein, 2020), while sex-disaggregated data for COVID-19

* Corresponding author.
E-mail address: nataliadiaz@ugr.es (N. Díaz-Rodríguez).
[1] Equal Contribution.
[2] Work done prior to joining AWS.

https://doi.org/10.1016/j.ipm.2023.103276
Received 27 April 2022; Received in revised form 8 December 2022; Accepted 9 January 2023
Available online 12 January 2023

in several European countries shows a similar number of cases between sexes, but more severe outcomes in aged men (Gebhard et al., 2020).

Biological differences in the immune system in men and women may affect the person's ability to fight COVID-19. It may be argued that men are more vulnerable to COVID-19 in relation to women because of a distinctive lifestyle, smoking, drinking, working hours, sex hormones, hypertension, and other circumstances (Smith, 2020). Research suggests sex-based differences in ACE2 and TMPRSS2 enzymes and the link between circulating ACE2 and COVID-19 (Gebhard et al., 2020) is not clear. Additionally, sex and gender may intersect with age and race, to further increase the risk of severe COVID-19 outcomes in men. In *PLoS pathogens* and *CMAJ* journals it is also discussed how other socio-economic factors also increase the risk of COVID-19 (Klein et al., 2020; Tadiri et al., 0000). Systemic health and social inequities have disproportionately exposed low-income communities, racial and ethnic minorities to higher risk of COVID-19 infection and death. Additionally, uneven testing strategies across the world, and the quality of epidemiological big data, limit the accuracy of estimated distribution of COVID-19 patients according to Kopel et al. (2020). The contributing factors can broadly be categorized into physical, sex-related attributes, lifestyle gender related attributes, and cultural, gender role related variables. The contribution, consistency of the effect and causal role of each of those groups of variables is very different and must be taken into consideration when building machine learning models, performing data analysis or making data-informed decisions. While physical, sex-based factors can be viewed as relatively constant predictors, the gender lifestyle attributes are fluid and vary from individual to individual. Furthermore, the cultural, gender roles based variables are intrinsically contextual and culture specific. Failing to adapt the model to cater for those individual and cultural differences hinges both the accuracy of the predictions and put the group of individuals under a threat of disparate impact of such predictions. However, the complex structure of the various factors that influence the disease may not be evident from the accessible health databases. Observational datasets coming from public information access systems can be fragmented, coming from diverse sources and may not necessarily include all the attributes relevant for the analysis. This paper showcases potential risks of biased or incomplete data and how causality can be put into practice as part of a risk management strategy to avoid discriminating systems. In this paper we focus on analysing the difference in causal and fairness impacts of different categories of variables linking sex or gender and COVID-19 severity. We demonstrate how omitting causal, research-based knowledge from the model of sex and COVID-19 relationships can further propagate more intricate forms of bias in computational models and lead to discriminatory and harmful pandemic policies and decision making.

As a result, we bring light into: (1) a potential set of hypotheses within our COVID-19 case study to further verify its causal link, and (2) the unintended consequences that can derive from a lack of an adequate toolbox to support fair and accurate decisions.

The contributions of this paper are the following:

- A review of the most up-to-date literature mainly from 2020–2022 analysing studies, from the gender and sex perspective, that indicate an increased male vulnerability to COVID-19;
- An identification of a set of hypotheses that can potentially be responsible for the observed disparities;
- Illustrating the causal relationship of sex/gender and COVID-19 with causal graphs detailing different groups of mediating or confounding factors;
- Performing causal analysis on synthetic data based on our model, to demonstrate the difference in effect of those factors and extent of discrimination they may cause;
- We highlight explainability and causality instrumental approaches to better understand big data and facilitate equitable data-driven decisions;
- A synthesis of our findings is contributed in the form of a *List of Confounders and Mediators*, which reviews the minimal ingredients necessary to account for when mitigating potential sources of bias possibly explaining reported disparities.

The rest of this paper is organized as follows. First we present the most recent literature on gendered and sex-related effects of COVID-19 in Section 2. In Section 3 we conducted a data analysis on publicly available big datasets to investigate the possible impact of gender-lifestyle-related confounding factors on the COVID-19 outcomes. Section 5 presents a prioritized check-list for identifying and including confounding and mediating variables into the data. Next we introduce the necessary preliminaries for discussing causal fairness frameworks and mediation analyses 4. We provide a synthetic data based causal analysis illustrating the necessity to consider confounders and mediators to avoid discrimination in Section 6. Finally, we discuss results and open research directions for the future in Sections 7 and 8 respectively.

2. Related work: Identifying causal explaining factors on sex/gender and COVID-19 relationship from epidemiological and clinical studies

We review findings based on big data on gender and COVID-19 from two angles. First, we analyse a body of papers placing gender and sex as a risk factor towards COVID-19, focusing on explaining the reasons behind disparity. Second, we categorize the explaining variables into mediators and confounders and discuss possible fairness implications of the former results that could lead to discrimination decisions, with the aim of guiding the causal design of the underlying model.

The amount of literature providing evidence on links between sex/gender and COVID-19 vulnerability is significant (Besserve, Buchholz, & Schölkopf, 2021). Table 1 shows articles finding men to be more vulnerable to COVID-19 in comparison to women. The explanations for this association are as well diverse. One of the possible factors is sex impact on vaccine acceptance, responses, and outcomes (Gebhard et al., 2020). Women are often less likely to accept vaccines but once vaccinated, develop higher antibody responses (Klein, Jedlicka, & Pekosz, 2010). For example, after vaccination against influenza, yellow fever, rubella, mumps, measles,

small pox, hepatitis A and B and dengue viruses, protective antibody responses are twice as high in adult females compared with males. However they report more adverse reactions to vaccines than males (Gebhard et al., 2020). Moreover, biological differences in the immune systems of men and women exist, and they may affect the capacity to fight COVID-19 infection. Men appear to be at a greater risk with COVID-19 compared to women, whose higher immunologic response is probably associated with decreased mortality (Chiarella, Pabelick, & Prakash, 2021). Furthermore, certain differences in cardiac manifestations in COVID-19 must be considered as a core component (Sharma, Volgman, & Michos, 2020). From the observational studies perspective, men appear to be at a greater risk. Sex is surely not the only risk factor in a disease that, according to Smith (2020), is challenging to diagnose and theorize, and whose effects also depend on vulnerabilities related to diabetes, obesity, hypertension, heart disease, chronic kidney disease, and chronic pulmonary disease according to Klein et al. (2020). Many authors suggest that women naturally produce more types of interferon, which limits the abnormal immune response in the form of serious cases of COVID-19. Moreover, women also produce more *T* lymphocytes which kill infected cells; and the "female" hormone estradiol would also offer greater protection against infection. On the contrary, studies indicate testosterone would limit the immune response in men, which may explain the observed sex-bias (Peckham et al., 2020; Traish & Morgentaler, 2021).

Immunitary response duration was studied at the Pasteur Institut[3] and CHU of Strasbourg on 308 healthcare personnel that developed a light form of COVID-19 (Grzelak et al., 2020). They show significantly steeper, i.e., faster decline in antibodies (anti-S and NAbs) in males than in females independently of age and BMI, hinting to a lower duration of protection after SARS-CoV-2 infection or vaccination. As more protective antibodies are formed in women, they last longer and so, women are better protected.

The relevance of gender norms, roles, and relations that influence women and men differential vulnerability to infection, exposure to pathogens, treatment received, as well as how these may differ among different groups of women and men is outlined in Wenham, Smith, and Morgan (2020).

When comparing the COVID-19 case fatality rate (CFR) between China and Italy, the authors in von Kügelgen, Gresele, and Schölkopf (2020) infer how methods from causal inference –in particular, mediation analysis–, can be used to resolve apparent statistical paradoxes and other various causal questions from data regarding the current pandemic. Many research studies (Head et al., 2020) revealed that systemic health and social inequities have disproportionately increased the risk of COVID-19 infection and death among low-income communities and racial and ethnic minorities. The outcomes in Bertsimas et al. (2020) provide insights on the clinical aspects of the disease, on patients' infection and mortality risks, on the dynamics of the pandemic, and on the levels that policymakers and healthcare providers can use to alleviate its toll. In the gender and social norms side, a recent study conducted in Spain (one of the hardest hit countries in Europe) reported that women had more responsible attitude towards the COVID-19 pandemic than men (De La Vega, Barquín, Boros, & Szabo, 2020), and another in the US showed that women take more precautions, wear more masks and cover more coughs than men.[4] Gender roles are considered as those influencing women's and men's different vulnerability to infection and exposure to pathogens, as reported in Wenham et al. (2020). The impact of gender-specific lifestyle, health behaviour, psychological stress, and socioeconomic conditions on COVID-19 is further studied in Gebhard et al. (2020).

According to most of the literature observed, the sex bias observed in COVID-19 as stated by Peckham et al. (2020), is a worldwide phenomenon suggested by observational and clinical research. However, the explaining factors listed in the discussed studies are diverse and non-uniform in their sensitivity to individual or cultural contexts.

3. Gender-related lifestyle habits and COVID-19 vulnerability

To help understand the association between gender and COVID-19, we conducted a more focused data analysis based on publicly available data to investigate the possible impact of sex and gender on the COVID-19 epidemic.[5] Even if ecological analysis[6] is considered the lowest form of epidemiological evidence, and potentially involves confounding variables, we are aware that it may not be a more accurate assessment than the individual level studies being surveyed in this article. Nonetheless, in this section we use this kind of analysis in order to elucidate plausible risk factors and unaccounted variables potentially explaining the disproportionate results.

We constructed a database that aggregates confirmed cases statistics, COVID-19 deaths, ICU admissions and smoking data per gender for 61 countries spanning 5 continents. The data sources are briefly described below.

- The *Global Health 50/50*[7] project housed at University College of London, which is created by a live tracker that aggregates data on COVID-19 cases and mortality from published government reports. At the time of our analysis on April 05, 2021, sex-disaggregated data for 183 countries including confirmed cases, confirmed deaths, etc. was represented in the live tracker.
- We also used a public dataset maintained by *Our World in Data*,[8] which also contains additional information such as smoking, population, and daily COVID-19 cases.

[3] https://www.pasteur.fr/fr/espace-presse/documents-presse/COVID-19-duree-reponse-immunitaire-neutralisante-plus-longue-femmes-que-hommes

[4] https://hbswk.hbs.edu/item/the-covid-gender-gap-why-fewer-women-are-dying

[5] Data analysis notebook in R available for reproducibility online: https://rpubs.com/wafaeB/684506

[6] Studies where individual features and outcomes are aggregated at a group level and then analysed.

[7] Global Health 50/50 project website https://globalhealth5050.org/.

[8] Our World in Data portal .

Table 1
Summary of claims involving statements regarding men being more affected by the COVID-19 compared to women. X indicates correlation of that variable with the COVID-19. M: men are more affected, F: women are more affected by COVID-19, SSD: Statistically Significant Difference, NSD: Non Statistically-significant difference. Factors: S: Smoking, D: Drinking, C: Cancer, H: Hypertension, DM: Diabetes mellitus, CD: Cardiovascular diseases, CRD: Chronic respiratory disease, CLD-chronic lung disease, HD: Heart disease, O: Obesity, II: Inflammatory immune responses, CHK: Chronic kidney disease, CPD: Chronic pulmonary disease. Even though most articles claim men are more affected by COVID-19 than women and die more, none of them shows statistical significance nor has enough data to provide causal links beyond correlational studies.

Study	Tested hypothesis	Men are more vulnerable	Reported health conditions	Age correlation	Reported drinking/ smoking
Impact of sex and gender on COVID-19 outcomes in Europe (Gebhard et al., 2020)	COVID-19 is deadlier for infected men than women	✓(NSD: M)	C, H, DM, CD, CRD, CLD	✓	✓(D, S)
Coronavirus: why men are more vulnerable to COVID-19 than women? (Bwire, 2020)	There are higher morbidity and mortality rates in males than females	✓(NSD: M)	O, DM, H	✓	
Biological sex impacts COVID-19 outcomes (Klein et al., 2020)	Mechanistic differences including the expression and activity of ACE2 enzyme result in antiviral immunity, cases, hospitalizations and deaths differences.	✓(NSD: M)	CPD, CKD, II, HD, O	✓	
COVID-19: the gendered impacts of the outbreaks (Wenham et al., 2020)	Men are more likely to remain hospitalized an die and less likely to be discharged from the hospital than women.	✓(NSD: M)	H	✓	✓(S)
Racial and gender based differences in COVID-19 (Kopel et al., 2020)	Ethnic differences influence susceptibility and mortality	✓(NSD: M)	HD, O, CLD, C, H, DM, CD	✓	✓(D, S)
Sex Differences in Mortality From COVID-19 Pandemic: Are Men Vulnerable and Women Protected? (Sharma et al., 2020)	Male sex plays a role in increased mortality rates	✓(NSD: M)	H, DM, CD, CRD, CLD	✓	
The influence of sex and gender domains on COVID-19 cases and mortality (Tadiri et al., 0000)	Gender Inequality Index is positively associated with male:female cases ratio	✓(SSD: M)	19		
Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission (Peckham et al., 2020)	Male sex is a risk factor for death and ITU admission but not for infections.	✓(SSD: M)	H, II, C	✓	✓(S)

By aggregating data from these two sources, and including only countries for which confirmed cases, deaths and smoking information is available. It is worth noting that, in this analysis, due to missing data for some countries, and taking into account the low granularity of the data, our choice was to focus only on the countries where all data columns were complete. We were able to analyse complete data from 89 countries.[9]

We then looked at the *male-to-female* (male/female) ratio of confirmed cases, ρ_{cases}, and compared it to the *male-to-female* ratio of deaths, ρ_{deaths}, for each country. We particularly classified countries on 4 groups based on these two parameters as follows (Table 2):

- Group 1 includes the countries in which $\rho_{cases} > 1$ and $\rho_{deaths} < 1$. In our analysis, only two countries belong to Group 1.
- Group 2, which contains 32 countries, represents countries in which $\rho_{cases} < 1$ and $\rho_{deaths} > 1$.
- Group 3 contains 7 countries in which $\rho_{cases} < 1$ and $\rho_{deaths} < 1$.
- Group 4 includes 49 countries in which $\rho_{cases} > 1$ and $\rho_{deaths} > 1$.

Among the analysed countries in our study, only Lebanon and Uganda belong to Group 1. Our analysis revealed that while there are more confirmed cases among men compared to women, i.e. $\rho_{cases} = 1.45$ in Lebanon and $\rho_{cases} = 2.18$ in Uganda, the male-to-female ratio of deaths is still smaller, i.e. $\rho_{deaths} = 0.44$ and $\rho_{deaths} = 0.86$ in Lebanon and Uganda, respectively. Therefore, more deaths were reported among women. Thus, this case seems to be contradictory to global data which indicates that men are more likely to get severely affected by COVID-19, and die more from the disease than women. One of the possible reasons is women's representation

[9] The total aggregated multi-source data contained the following countries: Albania, Tunisia, Mozambique, Montenegro, Cyprus, Bosnia and Herzegovina, Spain, Turkey, Romania, Netherlands, Argentina, France, Portugal, Switzerland, Iceland, Kyrgyzstan, Sweden, Poland, Latvia, Eswatini, Jamaica, New Zealand, Croatia, Cambodia, Armenia, Ukraine, Slovakia, Belgium, South Africa, South Korea, Canada, Hungary, Vietnam, Slovenia, Mongolia, Lithuania, Estonia, Bahamas, Qatar, Thailand, Malawi, Burkina Faso, Bangladesh, India, Pakistan, Nepal, Nigeria, Yemen, Congo, Oman, Kenya, Panama, Costa Rica, Singapore, Dominican Republic, Liberia, Myanmar, Morocco, Bahrain, Haiti, Mexico, China, Greece, Philippines, Maldives, Paraguay, Zimbabwe, Colombia, Denmark, Italy, Barbados, Sri Lanka, Ecuador, Malta, Iran, Rwanda, Finland, Brazil, Indonesia, Israel, Austria, Chile, Norway, Luxembourg, Germany, Australia, Lebabnon, Uganda.

Table 2
Summary of analysed male-to-female cases ratio and male-to-female deaths ratio.

More Deaths: *More Cases:*	Females	Males
Females	Group 3	Group 2
Males	Group 1	Group 4

in certain sectors strongly hit by the pandemic, such as the garment and textile sector, in some Asian and African countries. This can translate into two potential explanations motivating more deaths in women: (1) they become unemployed and without access to healthcare to deal with the disease, or (2) they become more vulnerable and most affected by cotton industry-related respiratory diseases related with the lack of safety equipment in unhygienic, unsafe environments with hazardous work conditions, as reported in Kabir, Maple, Usher, and Islam (2019), Silpasuwan, Prayomyong, Sujitrat, and Suwan-Ampai (2016). However, more research needs to be done in order to provide more insights on the vulnerability of women to COVID-19 in Vietnam.

In Group 2, which contains 32 countries, women were more contaminated by COVID-19 than men. However, the number of deaths among male was higher. Data for this group also shows that this might be related to the much higher smoking rate in these countries. As shown for example in Fig. 1, a very high male-to-female smoking ratios are observed in most of the countries in this Group. Particularly, the highest smoking rates are observed in Tunisia, Albania and Mozambique, which also have the highest smoking ratios. Note that, in Figs. 1, 2, 3 we applied log scaling to the calculated ratios in order to plot them on a comparable scale. That is, a positive male-to-female smoking or death log-scaled ratio indicates a higher number of smoking or death among men, while a negative male-to-female smoking or death log-scaled ratio indicates a higher number of smoking or death among women. While smoking might be one of the reasons that increases the risk of hospitalization and death by COVID-19, as it is the case for most respiratory diseases, more data is needed in order to provide evidence on this hypothesis, such as age, number of tests by gender, etc.

Driven by the observations we made in the previous group of countries, we were also interested in investigating the association between deaths ratios and smoking ratios for Group 3 and 4. Figs. 2 and 3 report the *male-to-female* deaths ratio vs the *male-to-female* smoking ratio for Group 3 and Group 4, respectively. Group 3 represents 7 countries in which both confirmed and fatality rates are higher for women compared to men, while Group 4 represents 49 countries in which both confirmed cases and deaths are higher for men. Figs. 2 and 3 also show a possible association between smoking and deaths. While the average *male-to-female* log-scaled smoking ratios is 1.5 across countries in Group 3, its value is higher and is up to 1.9 in Group 4, in which the *male-to-female* deaths ratios are also higher. It is also possible that countries in Group 3 are more likely to apply fairer testing strategies compared to the countries in Group 4, that have the highest *male-to-female* death ratios.

While our analysis suggests a possible association between smoking and a higher number of COVID-19 deaths, as most countries having a high *male-to-female* deaths ratios, have a high *male-to-female* smoking ratios as well, there is no firm conclusion that can be drawn regarding the relationship between smoking, sex and COVID-19. In addition, countries considered different criteria during the pandemic for reporting COVID-19 deaths and this could make understanding the impact of sex on COVID-19 ambiguous.

The differential findings and disparities observed across the four groups of countries in our analysis emphasize the need to understand why COVID-19 impacts some groups more than others. This might reflect other related factors and issues that need to be addressed, such as incomplete data and decision making biases. In the next sections we attempt to summarize the potential explaining variables found in the literature review and our analysis to structure it according to the ascribed role in the causal sex-COVID-19 relation framework.

4. Background on causality and mediation analysis

Variables are denoted by capital letters (e.g. X, Y). Small letters denote specific values of variables (e.g., $A = a$, $W = w$). Bold capital (e.g. $\mathbf{V}$) and small letters (e.g. $\mathbf{v}$) denote a set of variables and a set of values, respectively.

A causal graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, composed of a set of variables/vertices $\mathbf{V}$ and a set of edges $\mathcal{E}$, is a directed acyclic graph (DAG) that describes the causal relations between variables. Edges have causal interpretations. That is, a directed edge $X \rightarrow Y$ indicates a causal relation from the cause variable X to the effect variable Y. Consequently, if all other variables are fixed to some values and we change the value of X, Y will change, but not the other way around (changing the value of Y will not change the value of X).

There are three basic structures in a causal graph, namely, a mediator, a confounder, and a collider (Pearl, 2009). Fig. 4 shows an example of each one of these structures. Variable W in Fig. 4(a) is called a mediator because it mediates the causal effect of X on Y.[10] A confounder variable (C is a common cause of two other variables (X and Y). It is important to mention that in both mediator and confound structures, X and Y are correlated. The difference is that in a mediator, X is a cause of Y, but in a confounder, X is not a cause of Y. They are simply correlated. A collider, on the other hand, is a variable caused by two other variables (Z in Fig. 4(c)).[11] Unlike the two other structures, in presence of a collider, X and Y are not correlated. However, if we condition on Z, X and Y become correlated.

[10] The mediator structure is known also as chain structure.

[11] A collider structure is known also as v-structure.

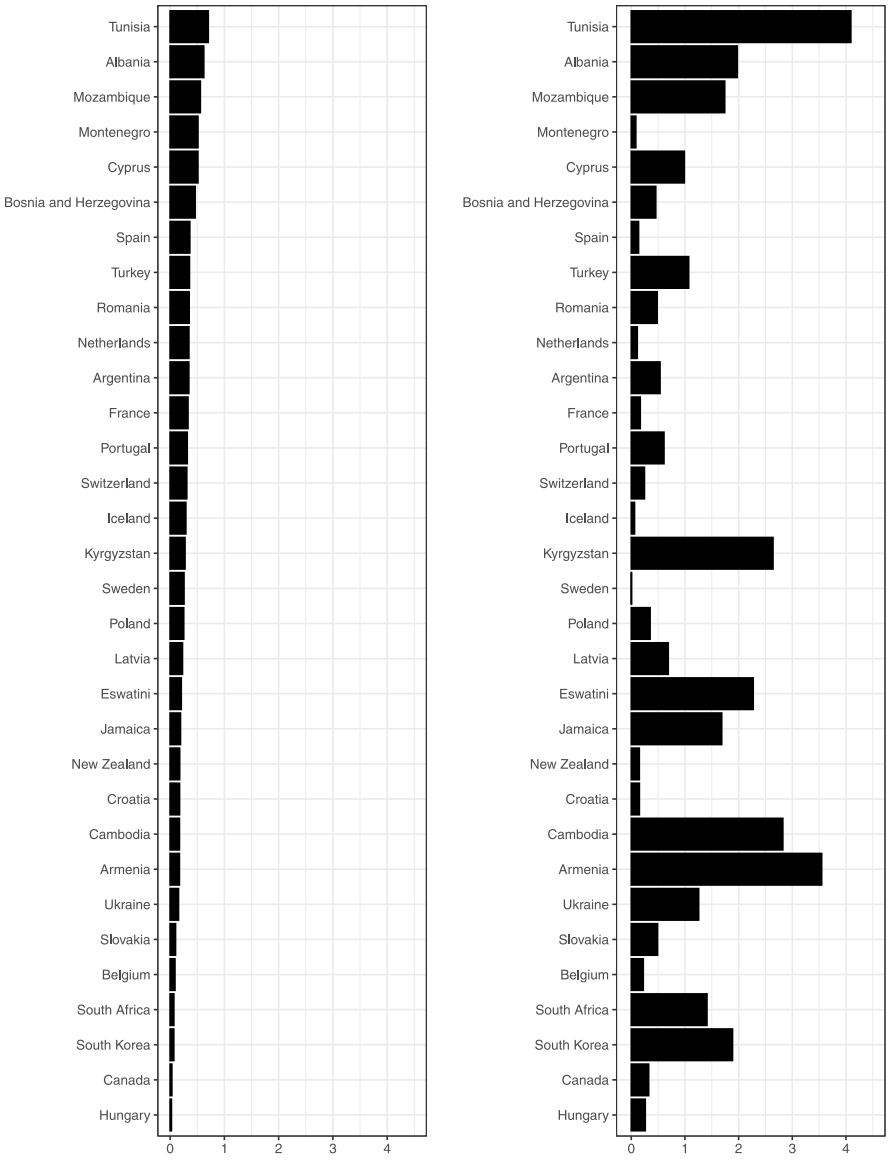


Fig. 1. Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* smoking -female-to-male- ratio smoking (Right) for Group 2 (more cases in women but more deaths for men). This group is composed by 32 countries and shows that one possible explanatory variable is the factor *smoking*, since men are shown to smoke more in these countries.

As the causal relation between two variables can go through different paths, mediation analysis consists in distinguishing these causal path-ways. For example, in Fig. 5, a causal effect between X and Y can be split into direct ($X \rightarrow Y$), indirect ($X \rightarrow R \rightarrow Y$ and $X \rightarrow E \rightarrow Y$), or path -specific effect (only $X \rightarrow E \rightarrow Y$). Assuming X is a sensitive variable (used for discrimination), this is very relevant to fairness as a direct effect is always unfair because the sensitive variable should not be used directly to decide about the outcome, while an indirect or path-specific effects may be unfair or fair depending on the mediator variable: an indirect effect through a redlining/proxy variable (R) is unfair, while an indirect effect through an explaining variable (E) is acceptable (fair). A proxy variable is a descendent of X which is significantly correlated with it in such a way that using the proxy in the outcome Y has almost the same impact as using X directly. An explaining variable is also a descendent of X used to decide about the outcome Y that is influenced by X in a manner that is accepted as nondiscriminatory. For example, a discrimination against women for job hiring is acceptable if it is justified by the low education level of female candidates. Deciding if a mediator is a proxy or explaining variable requires typically some expertise about the context of the problem.

Using causality allows to appropriately assess fairness (and consequently discrimination) due to two main reasons. First, by identifying confounder variables between X and Y, it becomes possible to account for the non-causal effect that goes through the confounder variables. For example, the effect going through the path $X \leftarrow C \rightarrow Y$ in Fig. 4(b) is non-causal while all paths between

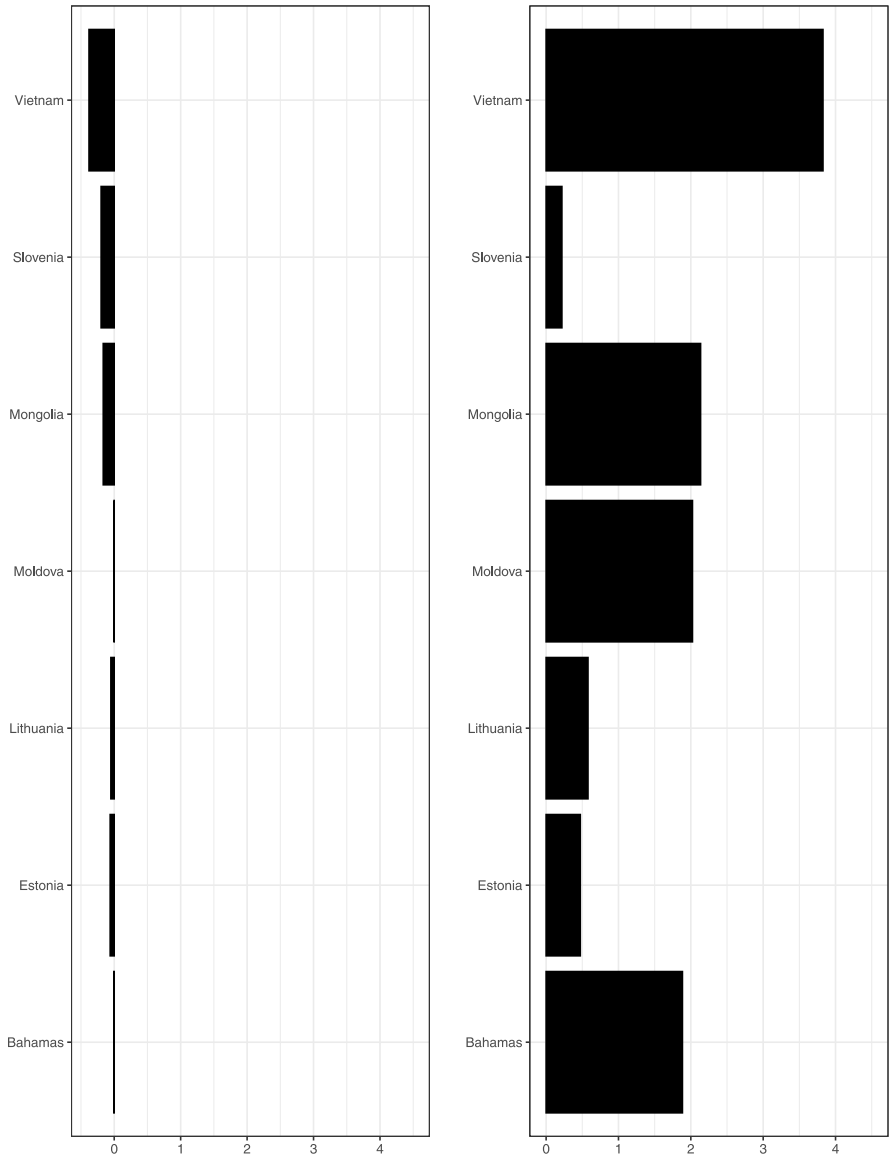


Fig. 2. Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* Smoking ratio (Right) for Group 3 (7 countries, in which both cases and death ratios are higher for women, i.e., the opposite of most articles claims). In these countries, women smoke almost equally as men, and thus, smoking does not seem to clearly be an explanatory variable: women die as much or more than men.

X and Y in Fig. 5 correspond to causal effects. This is the reason we say that 'causation is different than correlation'. Second, causal mediation analysis allows to split the total causal effect of X to Y into direct/indirect and fair/discriminatory effects.

The most common non-causal fairness notion is total variation (TV), known as statistical parity, demographic parity, or risk difference. The total variation of $X = x_1$ on the outcome $Y = y$ with reference $X = x_0$ is defined using conditional probabilities as follows:

$$TV_{x_1,x_0}(y) = P(y \mid x_1) - P(y \mid x_0) \tag{1}$$

Intuitively, $TV_{x_1,x_0}(y)$ measures the difference between the conditional distributions of Y when we (passively) observe A changing from x_0 to x_1.

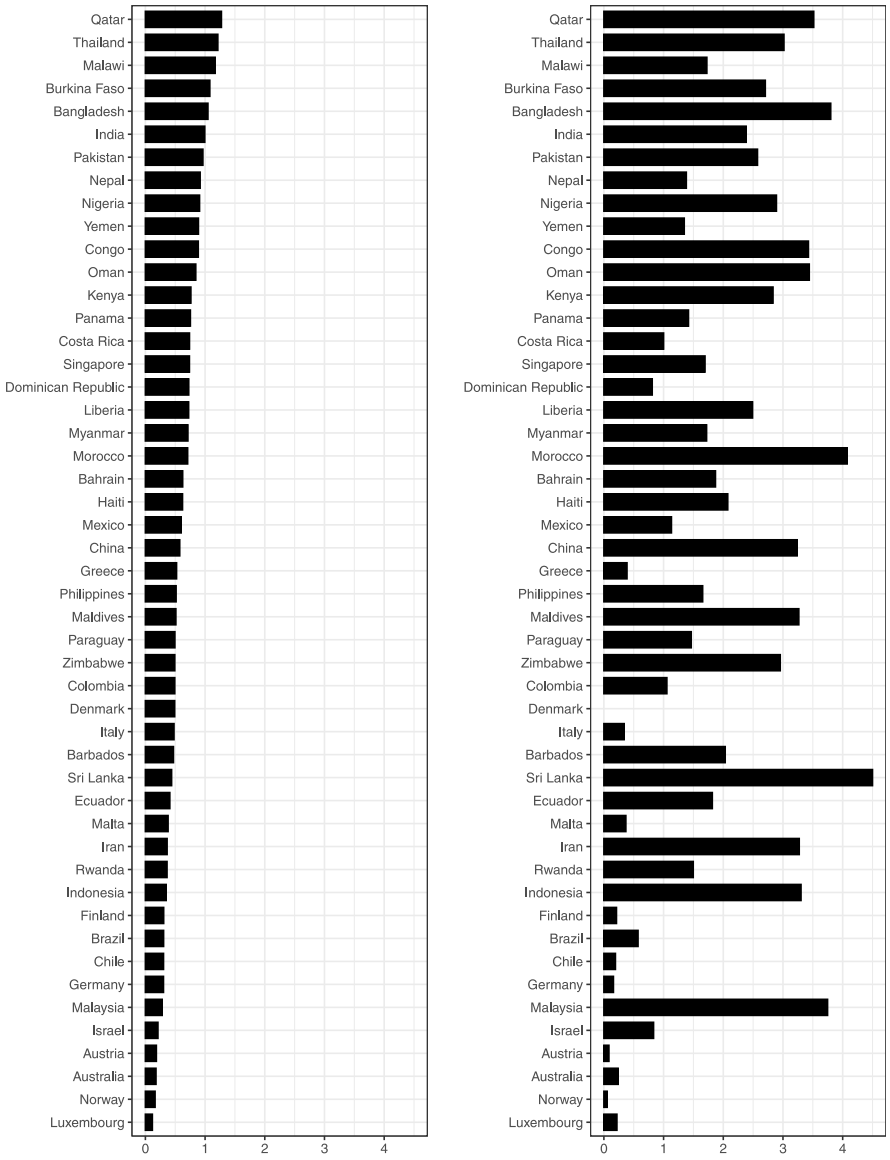


Fig. 3. Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* Smoking ratio (Right) for Group 4 (49 countries where both cases and deaths are higher for men). This plot may reveal different testing strategies, as men are always more impacted.

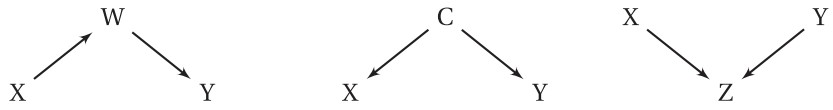
Total effect (TE) (Pearl, 2009)[12] is the causal version of TV and is defined in terms of experimental probabilities as follows:

$$TE_{x_1,x_0}(y) = P(Y = y|do(X = x_1)) - P(Y = y|do(X = x_0)) \tag{2}$$

TE measures the effect of the change of X from x_1 to x_0 on $Y = y$ along all the causal paths from X to Y. Intuitively, while TV reflects the difference in proportions of $Y = y$ in the current cohort, TE reflects the difference in proportions of $Y = y$ in the entire population. $P(Y = y|do(X = x))$ denotes the probability of $Y = y$ after an intervention $do(X = x)$. This is equivalent to probability of $Y = y$ after forcing all individuals in the population to have value $X = x$. $P(Y = y|do(X = x))$ is denoted $P(y_x)$ for short[13].

[12] Total Effect is also known also as average causal effect (ACE).

[13] The notations $Y_{X \leftarrow x}$ and $Y(x)$ are used in the literature as well. $P(Y = y|do(X = x)) = P(Y_{X=x} = y) = P(Y_x = y) = P(y_x)$ is used to define the causal effect of X on Y.



(a) Mediator (b) Confounder (c) Collider

Fig. 4. Basic structures of causal graphs.

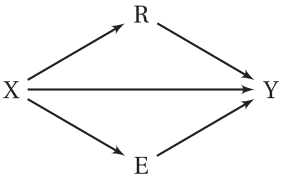


Fig. 5. The causal effect between X and Y can be split into three different paths: direct ($X \rightarrow Y$) and indirect ($X \rightarrow R \rightarrow Y$ and $X \rightarrow E \rightarrow Y$), involving (R)edlining/proxy and (E)xplanatory variables.

4.1. Mediation analysis to analyse causal effects

Mediation analysis is about distinguishing the different paths of the causal effect between two variables X and Y. Causal paths can be either direct or indirect. Natural direct effect (NDE) (Pearl, 2001) is the simplest notion of mediation analysis which measures the direct causal effect between two variables. (e.g. X and Y).

Assuming variable X is binary (it can take two possible values x_0 and x_1), NDE is defined as:

$$NDE_{x_1,x_0}(y) = P(y_{x_1,\mathbf{Z}_{x_0}}) - P(y_{x_0}) \tag{3}$$

Where $\mathbf{Z}$ is the set of mediator variables and $P(y_{x_1,\mathbf{Z}_{x_0}})$ is the probability of $Y = y$ had X been x_1 and had $\mathbf{Z}$ been the value it would naturally take if $X = x_0$. Using the graph in Fig. 5, this means that X is set to x_1 in the single direct path $X \rightarrow Y$ (there is always only one direct path but several indirect paths between X and Y) and is set to x_0 in all other indirect paths ($X \rightarrow R \rightarrow Y$ and $X \rightarrow E \rightarrow Y$).

Natural indirect effect (NIE) (Pearl, 2001) measures the indirect effect of X on Y and is defined as:

$$NIE_{x_1,x_0}(y) = P(y_{x_0,\mathbf{Z}_{x_1}}) - P(y_{x_0}) \tag{4}$$

Using the same graph (Fig. 5), this means that X is set to x_0 in the single direct path $X \rightarrow Y$ and is set to x_1 in all other indirect paths. The problem with *NIE* is that it does not distinguish between the fair (explainable) and unfair (indirect discrimination) effects.

The path-specific effect (Chiappa, 2019; Pearl, 2009; Wu, Zhang, Wu, & Tong, 2019) is a more nuanced measure that characterizes the causal effect in terms of specific paths. Given a path set π, the π-specific effect is defined as:

$$PSE^{\pi}_{x_1,x_0}(y) = P(y_{x_1|_{\pi},x_0|_{\overline{\pi}}}) - P(y_{x_0}) \tag{5}$$

where $P(y_{x_1|_{\pi},x_0|_{\overline{\pi}}})$ is the probability of $Y = y$ in the counterfactual situation where the effect of X on Y with the intervention $do(X = x_1)$ is transmitted along π, while the effect of X on Y without the intervention (x_0) is transmitted along paths not in π (denoted by: $\overline{\pi}$). For instance, in the graph of Fig. 5, if $\pi = X \rightarrow E \rightarrow Y$, then $\overline{\pi}$ includes $X \rightarrow Y$ and $X \rightarrow R \rightarrow Y$.

4.2. Other notions of causal fairness

In addition to total effect (TE) and related mediation analysis notions (NDE, NIE, and PSE), causal notions of fairness include qualitative notions such as *no unresolved discrimination* (Kilbertus et al., 2017), *no proxy discrimination* (Kilbertus et al., 2017) and *counterfactual fairness* (Kusner, Loftus, Russell, & Silva, 2017).

No unresolved discrimination (Kilbertus et al., 2017) is a fairness notion that focuses on the indirect causal effects from the sensitive variable X to the outcome Y. No unresolved discrimination is satisfied when no directed path from A to Y is allowed, except via a resolving (explaining) variable E. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner that is accepted as nondiscriminatory.

Similarly to no unresolved discrimination, *no proxy discrimination* (Kilbertus et al., 2017) focuses on indirect discrimination. A causal graph exhibits potential proxy discrimination if there exists a path from the protected attribute X to the outcome Y that is blocked by a proxy/redlining variable R. It is called proxy because it is used to decide about the outcome Y while it is a descendent of X which is significantly correlated with it in such a way that using the proxy in the decision has almost the same impact as using X directly. An outcome variable Y exhibits no proxy discrimination if the equality:

$$P(Y \mid do(R = r)) = P(Y \mid do(R = r')) \quad \forall\, r, r' \in dom(R) \tag{6}$$

Table 3
Causal explaining variables between gender/sex and COVID-19 severity, classified into mediators and confounders. Mediators are the intermediate variables on the causal path from sensitive attribute to the outcome. A confounder is a variable with incoming arrows in the graph to both sensitive attribute and an outcome (a cause for both) and creates spurious non causal relationship between the two.

Variable and source	Class	Group	Comments
Hormones (Chiarella et al., 2021; Dana et al., 2020; Klein et al., 2020; Peckham et al., 2020; Traish & Morgentaler, 2021)	Mediator	Sex-related Bio Var	Male hormone testosterone is associated with increased vulnerability, whereas female hormones are believed to play a protecting role.
Immune response (Chiarella et al., 2021; Dana et al., 2020; Grzelak et al., 2020)	Mediator	Sex-related Bio Var	More protective antibodies are formed in women and they last longer.
Smoking and drinking (Bwire, 2020; Gebhard et al., 2020)	Mediator	Gender-related Lifestyle Var	Higher smoking and drinking rates among men induce lung injuries that affect COVID-19 vulnerability.
Stress (Gebhard et al., 2020)	Mediator	Gender-related Lifestyle Var	Men often are more exposed to stress at work.
Hazardous industry (Kabir et al., 2019; Silpasuwan et al., 2016)	Mediator	Gender-related Lifestyle Var	It is worth noting that in some Asian countries women constitute a majority of garment and textile sector workers that are exposed to unsafe work conditions and are reported to be hit by the pandemics more than men.
Health behaviour (De La Vega et al., 2020; Gebhard et al., 2020; Tadiri et al., 0000)	Confounder	Gender roles related Var	Women are more health -conscious and compliant with health recommendations
Exposure to pathogens (Wenham et al., 2020)	Confounder	Gender roles related Var	In traditional societies women stay at home, and therefore are less exposed to the virus.

holds for any potential proxy variable R.

The use of both no unresolved discrimination and no proxy discrimination in real scenarios is limited by the assumption of valid causal graph availability. Hence, both fairness notions depend on the correct output of the causal discovery task.

Counterfactual fairness (Kusner et al., 2017) is a very strong fairness notion that requires equality between the observed outcome and the counterfactual outcome for every individual. That is, an outcome Y is counterfactually fair if under any assignment of values $\mathbf{V} = \mathbf{v}$ and any individual in the population,

$$P(y_{x_1} \mid \mathbf{V} = \mathbf{v}, X = x_0) = P(y_{x_0} \mid \mathbf{V} = \mathbf{v}, X = x_0) \tag{7}$$

where $\mathbf{V}$ represents the set of all remaining variables (all variables in the causal graph except $\{X, Y\}$). Counterfactual fairness, as an individual fairness notion, is satisfied if the probability distribution of the outcome Y is the same in the actual and counterfactual worlds, for every possible individual. For an exhaustive list of causal-based fairness notions, we refer interested readers to the survey of Makhlouf et al. Makhlouf, Zhioua, and Palamidessi (2020).

5. Confounders and mediators between sex and COVID-19 vulnerability

In this section we summarize the variables linking sex or gender and COVID-19 vulnerability, and categorize them in mediators and confounders (Table 3). The mediators are further divided into constant (sex-related) and varying (gender-related) from individual to individual. It has to be noted, that this classification is dependent on the goal of the predictor, or a question that we are trying to answer. Which of the variables are considered confounders, as well as which mediators can be viewed as explaining variables (fair) or redlining (unfair) is context specific. For example, when predicting the probability of mortality from COVID-19 and allocating resources in the ICU (intensive care unit), the variable responsible for gender-related health-consciousness is a confounder. Namely, it does not directly influence the development of a disease in the hospital, but creates a spurious correlation in the epidemiological data. On the contrary, if the predictor was trying to answer the question which group should be more targeted by health-related social advertising (urge to wash hands or wear masks), the same variable could be used as an explaining mediator. Thus the men could be targeted more, proportional to a measured effect of the health-consciousness gender-related bias.

In Table 3, we consider COVID-19 severity and mortality risk as a prediction question, and healthcare resources allocation in the hospital as a decision based on the perceived level of severity. From observational analyses and tables in previous section we can observe a set of factors repeating as conditioning factors to explain the differences of sex and gender's impact on COVID-19 vulnerability. In this section we synthesize these factors to provide an overall aggregation of COVID-19-related claims most stated by the literature on the impact of different variables on COVID-19.

We are aware that other studies have considered other factors as important ones in the way COVID-19 infection translates into a severe case, for instance, the blood group type (Pourali et al., 2020; Zietz & Tatonetti, 2020; Zietz, Zucker, & Tatonetti, 2020), vitamin D deficit (Ebadi & Montano-Loza, 2020; Jain et al., 2020), or other genetic factors (Zeberg & Pääbo, 2020). However, here we address only gender or sex related factors and their roles in predicting COVID-19 vulnerability.

Next sections will elaborate on the causal tools available to further study and corroborate such causal hypotheses and explanatory factors drawing on the body of analysed literature.

6. Avoiding potential discriminating policies through a causal approach

In general, fair decision should not be based on any knowledge of the sensitive attribute such as gender, race, sexual orientation, etc. The case of medical diagnosis and treatment is an exception, because certain diseases and conditions are specific to a particular sex, for example breast cancer which is almost exclusively characteristic for females. However, it is important to evaluate the exact extent of how much a physical component of being a female reduces the risk of mortality rather than gender related mediators or confounders. Current COVID-19 research shows that the underlying causes of vulnerability are diverse and unequal in the causal quality. As a consequence, this opens many pathways for the results to be distorted. Specifically, an already widely accepted discovery of women being more resilient to the disease can be affected by a spurious confounder or mediator which is not necessarily present in all women, and must be considered individually. Next, we illustrate unintended negative consequences for women if clinicians or governments base decisions on assumptions of greater resilience to the virus for females without adjusting for individual and cultural differences. We demonstrate the urge for more fine-grained causal analysis by performing mediation analysis on synthetic data generated following the epidemiological research informed causal model.

6.1. Data generation and model

To illustrate different causal paths between gender and COVID-19 severity we construct a causal model based on the discussed literature. We note that causal models can also be learned from data directly with causal discovery methods such as (Goudet et al., 2018; Shimizu, Hoyer, Hyvärinen, Kerminen, & Jordan, 2006; Spirtes & Glymour, 1991). However, expert knowledge and previous research in the domain is important in informing what variables have to be included in the data. Furthermore, a recent study (Binkytė-Sadauskienė, Makhlouf, Pinzón, Zhioua, & Palamidessi, 2022) shows that different causal discovery algorithms may not always agree on the resulting causal structure, therefore a combination of prior causal knowledge (for example, from experimental research) and statistical methods can help to achieve more robust results.

In Fig. 6 we provide a Directed Acyclic Graph (DAG) to represent the causal structure of the data generating process. A DAG is a graphical representation of independence properties of joint probability distributions. It is constructed from the nodes that represent the variables and the edges that denote conditional probability relationships. In our case the joint probability of the variables in the DAG can be factorized as follows:

$$P(G, S, L, B, C) = P(G)P(S|G)P(L|S)P(B|S)P(C|L, B, G) \tag{8}$$

Where $P(G)$ is the probability of observing (different) Gender Roles (Equal or Traditional), $P(S|G)$ is the probability of entering the set of samples where $Sex = Female$ or $Sex = Male$ is observed within the infected patients given the value of Gender Roles, $P(L|S)$ is the probability of unhealthy lifestyle given Sex, $P(B|S)$ is the probability of biological factors (BioVar) serving as a protection against COVID-19 complications given Sex, and $P(C|B, L, G)$ is the probability of observing severe COVID-19 disease given BioVar, Lifestyle and Gender Roles. An important difference between a Causal Graph in Fig. 6, and a Bayesian Network or Markov Chain is that parents of an edge are indicated based on assumed causal relationships (Pearl & Dechter, 2013) For example, despite the symmetric conditional independence relationship between symptoms and a disease (it is possible to predict symptoms given the disease or disease given the symptoms) the symptoms cannot be denoted as a cause for a disease (in nature the disease, for example the infection, happens before the symptoms). As defined by Pearl (2009), one of the most important properties of a causal DAG is that all nodes are independent of their non-descendants given their (immediate cause) parents. For example, given the model in Fig. 6, COVID-19 Severity (S) becomes independent of Sex conditioned on all the intermediate parents such as BioVar (B), Lifestyle (L) and Gender (G) Roles. Formally:

$$P(C) \perp\!\!\!\perp P(S)|P(B, L, G) \tag{9}$$

The model indicates Sex as a ‘‘treatment variable’’ and COVID-19 as an ‘‘outcome’’ variable. That means we will estimate the effect of Sex on COVID-19 severity through various mediating and confounding paths. We must note that we include Sex, not Gender as a ‘‘treatment’’ variable, because Sex is usually included in the healthcare records. We distinguish aspects of cultural gender that are important for COVID-19 outcome as mediators and confounders and suggest that they should be explicitly taken into account when performing the analysis.

The model includes three groups of variables that serve as mediators or confounders between a sensitive attribute (sex) and the prediction (COVID-19 vulnerability) (Table 3). We consider *BioVar* as biological, sex-related attribute or a set of attributes such as differences in hormones, immune reactions and others, as one type of mediator variables that are constant (or almost constant) for biological *Sex*. The next group of mediators, *Lifestyle*, are related to sensitive attributes only through correlations between *Sex* and certain lifestyle choices such as smoking or drinking habits. Those attributes are gender-related. They can vary from individual to individual and cannot be automatically inferred from a *Sex* variable in the data.

Finally, the variable *Gender Roles*, account for spurious correlations between gender and COVID-19 severity and are considered confounders. This group is less intuitive to understand, because it is expressed with an incoming arrow from *Gender Roles* to the *Sex* variable, but *Sex*, as any sensitive variable, is considered to have temporal priority so it cannot be caused by other variables. However here we follow the Fairness Model by Zhang and Bareinboim (2018) and conceptualize *Gender Roles* variables not as causing *Sex* in

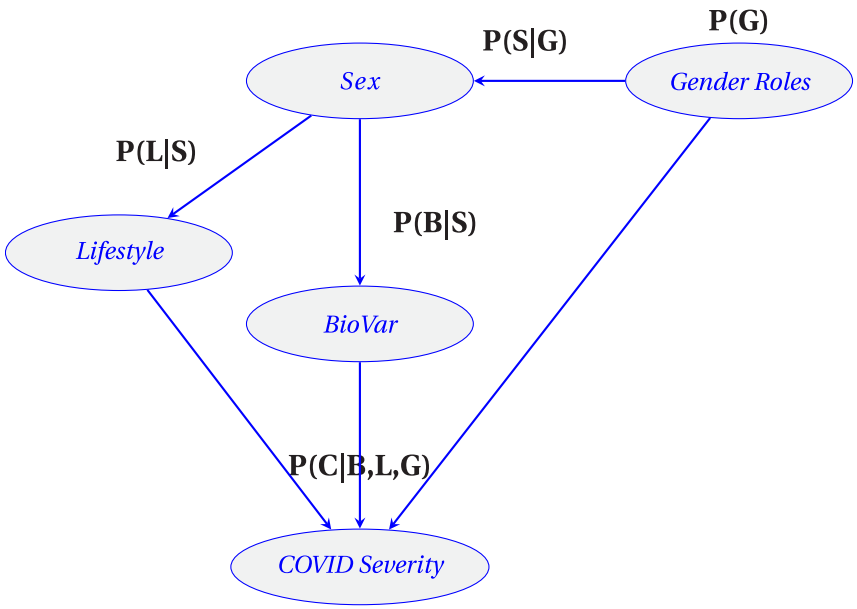


Fig. 6. The model for the variables explaining the link between sex/gender and COVID-19 severity informed by the gender/sex related COVID-19 research. **P(G)** - prior probability of observing a given Gender Role (Traditional or Equal). We make no assumptions and consider the probability to be equal for both values. However the prior probability is culture/country specific and can be informed from local sociological research. The prior and conditional probabilities for Bernoulli variables are explained in Eqs. (10)–(14). The Lifestyle and BioVar variables are considered mediators (on the causal path from Sex to COVID-19 Severity). The Gender Roles variable is considered a confounder (spurious effect between Sex and COVID-19 Severity) and reads as "the probability of observing each value of sex among hospitalized individuals and the probability of severe development of the disease for each value of sex depending on the gender roles". The graph accounts for the fact that culture-based gender roles may be causing a particular sex to behave in a certain way that affects the way he/she gets exposed to COVID-19, and exaggerates the effect of sex on developing severe COVID-19 disease. We thus set as confounder the variable Gender Roles (it is a back-door path because it points both at the cause -Sex and the effect -COVID-Severity). Thanks to mediating analysis we find the extent to which sex causes COVID-Severity, which results on an effect being non causal.

the real world, but as causing the proportion of certain *Sex* values *in the sample* or a sampling bias. For example, traditionally, women are viewed as more careful and compliant with healthcare recommendations. This reduces the risk of getting the disease and the development of the disease under domestic treatment conditions. This results in less female cases among hospitalized individuals. However, it being more cautious has no effect on the further development of the disease when the patient is already in the hospital and is taken care of by medical staff. We also include a variable Y in our model to express a policy or treatment decision based on the predicted severity of the disease. We discuss the implications of the different combinations of paths causing the outcome and particular decision in Section 6.3. Note that we build our model only to illustrate different paths between *Sex* and the *Severity* of the disease, but not to predict the actual severity in the individual. Therefore we do not include other variables important for the COVID-19 outcome not related to *Sex*, such as non-gender-related health conditions. Some of the variables could be related to race or social status, and a more complex model is required to account for them. However, it goes beyond the scope of this article and could be foreseen for future work.

The relationship between variables encoded in a DAG provides the means of recognizing conditional independence and identifying the set of parameters needed for any given computation (Geiger & Pearl, 1990). The model allows to identify the set of covariates for performing mediation analysis to evaluate the effect of Sex on COVID-19 Severity in the synthetic dataset (faithful to the model). The reason behind using synthetic data is twofold. First, the scenarios we are seeking to illustrate are related to the impact on individuals and require individual level data which is not freely available. Second, the purpose of this analysis is purely illustrative. Therefore the use of synthetic data and metrics derived from it help to convey the message without the danger of implying the usage of the results directly for clinical applications. The derivation of the real metrics for a particular use case has to come from quality individual level data that is representative of the local population and gender related cultural factors.

We generate binary data respecting the relations described in the model we built based on our review of gender/sex related COVID-19 literature and the groups of mediating/confounding variables we distinguish in Table 3. Fig. 6 illustrates the dependencies between the variables expressed as prior and conditional probabilities. Note that for the sake of simplicity of illustration the possible variables from each group are expressed as one combined group variable. The prior and conditional probabilities ($P(G)$, $P(S|G)$, $P(L|S)$, $P(B|S)$ and $P(C|B,L,G)$) we assign to the variables are not based on estimations from the data, but we respect the causal directions described in the literature. For example, we set the probability of the protective value of biological variables (BioVar) to be almost coinciding with the female sex (0.99%) and non-protective value with male sex.

Similarly, the probability of healthy Lifestyle is higher for females, but it is less deterministic than the biological factors (Bwire, 2020; Gebhard et al., 2020), Gender Roles give rise to lower probability to observe females in the data (women get hospitalized less,

perhaps because they take better precautions in daily life), as well as increased probability of mild rather than severe COVID-19 disease (Severity variable). Here we assume biological variables to have the largest overall effect on COVID-19 severity, lifestyle choices being the second, and gender role confounders as having the most moderate effect. The real proportions in the effect on COVID-19 severity can only be derived from a dataset including the relevant explaining variables for association between sex/gender and COVID-19 severity. We hope this article will encourage a causal analysis by proposing the model based on relevant research drawing attention to the relevance of causal knowledge for fair and explainable predictions.

The data is generated as follows. For simplicity all variables are set to be binary Bernoulli variables $\mathcal{B}$ with domain $k \in \{0, 1\}$ and parameters $0 \leq p \leq 1$ and $q = 1 - p$. The initial probability of the Gender Role variables being *Traditional* or *Equal* is set to be the same, namely 0.5 percent for each value. It can be adjusted based on our belief about a particular society where the data is collected.

$$GenderRoles \sim \mathcal{B}(0.5) \tag{10}$$

The Sex variable is set to be dependent on the Gender Roles variable. Namely, in the Traditional setting, women commute and are believed to be more health-conscious (De La Vega et al., 2020; Gebhard et al., 2020; Tadiri et al., 0000; Wenham et al., 2020), therefore, we observe overall smaller number of infected or severely ill female individuals. The conditional probabilities of Sex given Gender Roles reflect that hypothesis, but in absence of research on exact proportions, the numbers used are fictional. Under equal Gender Roles this effect is not observed, therefore the proportion of both sexes is equal.

$$Sex \sim (GenderRoles; p) = \begin{cases} Male : p_1 = 0.7, & if\ Gender\ Roles = Traditional, \\ Female : p_2 = 1 - p_1 & \\ Male : p_1 = 0.5, & if\ Gender\ Roles = Equal. \\ Female : p_2 = 1 - p_1 & \end{cases} \tag{11}$$

This means, that in case of Traditional Gender Roles the probability of getting infected (entering the sample) is much higher for men, whereas in the equal society the probability of getting sick is the same for both sexes.

The biological variables BioVar such as sex hormones or immune system specifics (Chiarella et al., 2021, 2021; Dana et al., 2020, 2020; Grzelak et al., 2020; Klein et al., 2020; Peckham et al., 2020; Traish & Morgentaler, 2021) are treated as almost deterministically dependent on Sex. We acknowledge, that more research on the individual fluctuations of those parameters would benefit the model.

$$BioVar \sim (Sex; p) = \begin{cases} Protective : p = 0.01 & if\ Sex = Male, \\ Protective : 1 - p & if\ Sex = Female. \end{cases} \tag{12}$$

Unhealthy lifestyle $value = 1$ (such as unhealthy lifestyle due to smoking, drinking, stress, etc.) are set to be more likely for men than for women (Bwire, 2020; Gebhard et al., 2020). The exact proportion is not grounded in the literature and is for illustration purposes only.

$$Lifestyle \sim (Sex; p) = \begin{cases} Unhealthy : p = 0.7 & if\ Sex = Male, \\ Unhealthy : 1 - p & if\ Sex = Female. \end{cases} \tag{13}$$

This would mean that probability to observe a male leading unhealthy lifestyle is 70% compared to only 30% probability of encountering a female with the same unhealthy habits.

Finally, we define probability of COVID-19 Severity as a linear combination of the previously discussed variables. The proportions of the impact of each group of variables in the equation is motivated by the corresponding volume of the research supporting the hypothesis in the reviewed literature at the time when this study is performed. Note that linearity of the effect is only an assumption made for simplicity and does not imply the real interaction between different factors.

$$COVIDSeverity \sim (Gender\ Roles, BioVar, Lifestyle; p) =$$

$$\begin{cases} Severe\ COVID - 19 : p = 0.2 \times Gender\ Roles + 0.5 \times BioVar + 0.3 \times Lifestyle \\ Mild\ COVID - 19 : 1 - p. \end{cases} \tag{14}$$

This means that the probability of severe COVID-19 disease is defined by the linear combination of the previously discussed variables.

However, the true functional form and exact proportions of the impact of each variable can be learned from the complete epidemiological data and is subject to future epidemiological research.

6.2. Mediation analysis to analyse causal effects of sex on the severity of COVID-19

To determine the proportion of the effect each of the variables has on the severity of the disease we perform causal mediation analysis (Fig. 8).[14] Similarly to von Kügelgen et al. (2020), where the proposed confounder is the age, we analyse the total effect

[14] The code can be found in the repository https://github.com/RuSaBin/Covid_Gender.

and the effect of the mediating variables under the confounding variables of *Gender Roles*. We apply causal fairness notions such as Total Effect (TE (2)), Natural Direct Effect (NDE (3)) and Path Specific Effects (PSEs (5)) through each mediator to determine the sex/gender bias in COVID-19 severity; namely, how much more likely is to observe a severe COVID-19 case for a man than for a woman.

To compute path-specific causal effects (PSEs) we use the imputation-based estimation of counterfactual outcomes implemented in R in the open-source Paths Library[15] (Zhou & Yamamoto, 2020) designed to trace causal paths from experimental and observational data. We use mediation analysis to estimate the proportion of the causal effect from Sex to COVID-19 Severity that is explained by one of the mediating variables. The imputation approach provides $K+1$ models that describe the expectations $\mathbb{E}[Y|\mathbf{X}, A], \mathbb{E}[Y|\mathbf{X}, A, M_1]...\mathbb{E}[Y|\mathbf{X}, A, M_k]$, where A is a sensitive attribute, $\mathbf{X}$ is a set of covariates, and $M_1...M_k$ are mediators (Makhlouf et al., 2020). For more extensive explanations of the Causal Fairness Notions we refer the reader to the survey of Makhlouf et al. (2020).

We also compute the Total Variation (TV, Eq. (1)). TV is a non causal fairness metric, and thus, it does not distinguish mediators from confounders. Note that in absence of confounders, TV and TE are equivalent. Hence, intuitively, the Confounding Effect.[16] (CE) can be estimated by subtracting the Total Effect from the Total Variation: $CE = TV - TE$[17] It is important to consider TV in our study as it corresponds to simple correlation between the Sex/Gender and the COVID-19 vulnerability. In contrast, the remaining metrics are more fine-grained in considering a specific path between Sex/Gender and COVID-19 vulnerability.

All causal effects are obtained by subtracting the probability of severe COVID-19 being a man from the same probability while being a woman. Hence, a positive value indicates men are more likely to develop severe COVID-19 case, while a negative value indicates a COVID-19 severity bias for women. A value of zero means that the probability of the outcome is equal for men and women. A value equal to one or minus one would indicate extreme cases, where the probability of severe COVID-19 disease is equal to one hundred percent for one group and zero for the other. The Confidence Intervals (CI) for each value are calculated via bootstrapping methods included in *paths* library and indicate the significance of the effect (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). All effects except Natural Direct Effect indicate a severity bias for men (Table 4). The Natural Direct Effect is negative, close to zero, and the corresponding CI includes zero[18] indicating that the detected effect between Sex and COVID-19 severity is not significant.

There is no Natural Direct Effect of Sex on COVID-19 because we assume that all the influence of sex/gender on COVID-19 severity is explained by the mediating variables (Table 3), even if some of them, such as BioVar (for example, female hormones) almost exactly correspond to sex. We make the decision to separate sex in general from specific sex related bio variables, to make it possible to account if needed, for individual fluctuations in those attributes and emphasize the more detailed explainability of the effect of sex on the disease severity. Following the interpretation of mediation analysis by Pearl (2014) we discover that being a male increases the overall risk (Total Effect Eq. (2)) of severe case of COVID-19 by 65.4%. Note that this is different from the 70% estimated by the Total Variation (Eq. (1)) before adjusting for confounding variables. Since the Direct Effect of Sex on COVID-19 severity is negligible (0.015, Eq. (3)) the causal effect that links Sex and COVID-19 severity is composed entirely of an indirect effect through BioVar of 52.1% and an indirect effect through lifestyle of 14.8%.

We illustrate the contrast between estimating Total Variation or performing Mediation Analysis in Fig. 7.

Let us say that we set a 10 h minimum amount of hours of medical attention for hospitalized COVID-19 patients as a baseline.. We want to allocate additional hours proportional to the risk of developing a severe COVID-19 outcome. In the case of computing Total Variation of effect of Sex on COVID-19 severity Fig. 7(b)) we would allocate male patients 70% ($TV = 0.7081$ in Table 4) more of time, namely 17 h. However, in the second case (Fig. 7(c)), assuming that sex is almost a perfect proxy for biological variables (BioVar), we allocate male patients only 50.2% more of resources, namely 15.2 h (given that women get 10 h). Additional attention hours are allocated to smoking patients, regardless of sex, proportionally to the effect of smoking on severe COVID-19 disease: 11.48 h for female smokers and 16.68 h for male smokers (we add additional hours to the minimum based on path specific effect through unhealthy lifestyle 0.148 in Table 4). Note that the synthetic data is generated assuming a conservative scenario, where the most significant part of the effect is due to biological variables which are closely correlated with sex (Table 3). If a larger part of the total effect was due to confounders such as gender roles related behaviour, the disparity between the Total Variation and Total Effect would further increase.

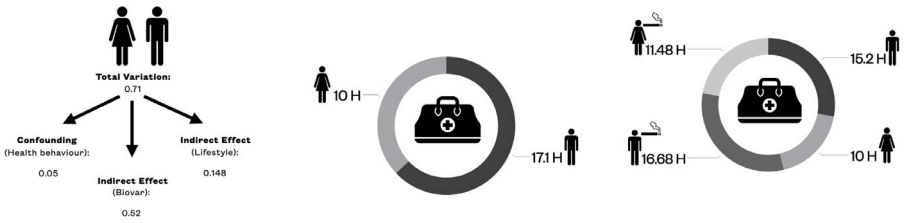
The amount of the effect caused by lifestyle or BioVar mediating variables, or the Gender Role confounding variable are not causally equivalent. In the following section we elaborate on the differences between BioVar or Lifestyle mediating variables, such as smoking or drinking habits and confounders responsible for spurious non causal effects, such as compliance with the healthcare recommendations (the variables and their belonging to the groups are listed in Table 3). We discuss the danger of failing to account for them in COVID-19 related policy making .

[15] https://github.com/cran/paths We direct the interested reader to the comprehensive survey on the libraries to perform mediation analysis in Starkopf, Andersen, Gerds, Torp-Pedersen, and Lange (2017).

[16] Also known as Spurious Effect

[17] This is not a formal definition, as the formula does not necessarily apply in non linear settings; however it is sufficient for illustrating the confounding effect in our data.

[18] which means that it is either small negative, small positive, or zero.



(a) The decomposition of the Total Variation of severe COVID-19 risk for males and females.

(b) The allocation of hours of medical care proportional to the estimate of Total Variation, namely if the minimum is 10h for women, the men are allocated 70 percent more - 17 hours.

(c) The allocation of hours of medical care proportional to the estimated path specific effects of Sex on COVID-19 Severity estimated by mediation analysis.

Fig. 7. The illustration of resource allocation according to different estimations of effect of Sex on COVID-19 severity via mediation analysis.

Table 4
Mediation Analysis of Causal Effects that illustrate the different paths of the influence of sex on COVID-19 severity. All effects except Direct Effect indicate a severity bias for men (positive values). The Direct Effect is close to zero, because we assume through the causal graph used as prior model of the world that all the influence of sex/gender on COVID-19 severity is explained by the mediating variables (either BioVar or Lifestyle variables) The effect caused by BioVar mediating variable is higher than the effect caused by the Lifestyle mediating variable. The last two columns of the table indicate lower and upper bounds for confidence intervals for the estimated effect values.

Variable	Estimated effect	Standard error	CI lower 95%	CI upper 95%
Natural Direct Effect:				
$Sex- > COVIDSeverity$	−0.015	0.035	−0.061	0.048
Path-Specific (Indirect) Effect :				
$Sex- > BioVar- > COVIDSeverity$	0.521	0.039	0.498	0.609
Path-Specific (Indirect) Effect:				
$Sex- > Lifestyle- > COVIDSeverity$	0.148	0.031	0.084	0.168
Total Effect:				
Sex $- > COVIDSeverity$	0.654	0.008	0.644	0.668
Total Variation:				
$Sex- > COVIDSeverity$	0.7081	–	–	–
Confounding Effect:				
$Sex < -GenderRoles- > COVIDSeverity$	0.0541	–	–	–

6.3. Disparate impact of sex on COVID-19 treatment decisions

Decisions in real life based on biased data can create disparities in treatment or disparate impact resulting in disadvantage for protected groups.

Disparate treatment is a variation in decisions for individuals that depends on the values of a sensitive attribute. *Disparate impact* occurs when decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017).

The adequate evaluation on fairness of decisions depends on the situation where the data analysis results will be applied. Here we would like to illustrate the evaluation of fairness in the COVID-19 pandemic context, by modelling a situation where inference about women being less vulnerable to the virus is used for assigning a priority treatment to an individual (for example, a longer hospitalization, closer monitoring or priority access to vaccines).

We assume that higher vulnerability or risk of severe symptoms for men is inferred from observing more cases of hospitalized individuals in the electronic health records data. We will illustrate different implications on fairness when classifying individuals based on *Sex* in combination with three groups of variables: lifestyle mediators (Lifestyle), biological mediators (BioVar) and gender-roles (Gender Roles) confounders In reality other variables not related to sex, as well as other health condition indicating features, can influence individual vulnerability to the virus. Thus, a thorough causal analysis becomes even more relevant: decisions should be based on known causes of vulnerability rather than on a sensitive attribute.

Scenario 1: *Disparate impact due to Gender Roles confounding variables*

In this case the confounding variable *Gender Roles* indicates whether the member of a sensitive group follows traditional or equal gender behaviour models. Under the traditional setting we assume that women are more careful and compliant than men, which makes them less likely to get COVID-19, as well as more likely to improve their condition when sick at home However, once in a hospital, where the patient is taken care of by the medical professionals, the impact of being more careful diminishes. Furthermore, individuals that do not follow the traditional gender-related behaviour might not fall into the same pattern. Failing to adjust the

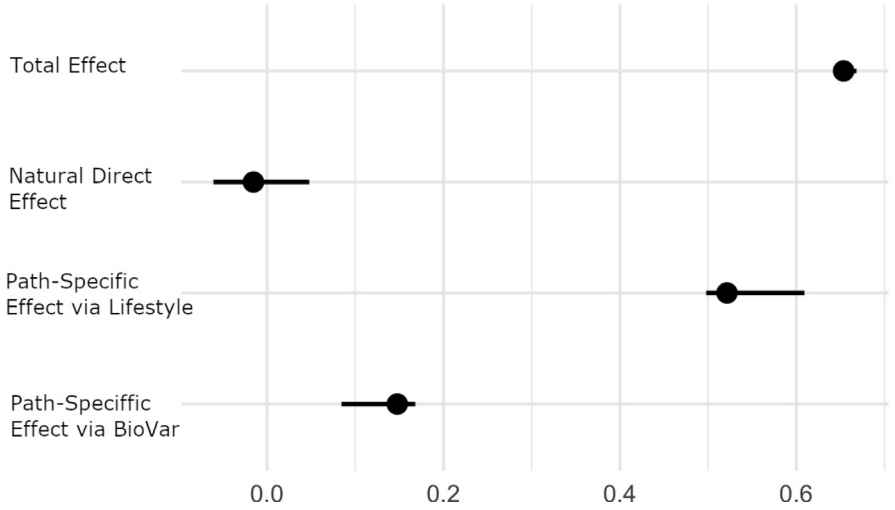


Fig. 8. The graphical summary of fairness notions, Total Effect (TE), Natural Direct Effect (NDE), and Indirect Effects (NIE) on COVID-19 severity through biological (BioVar) and Lifestyle variables along with their confidence intervals. These metrics indicate the difference in the probability of a severe form of COVID-19 disease for men and women. Positive values indicate that Sex = Male is associated with higher probability of severe COVID-19 disease than Sex = Female. A negative value for NDE would mean the opposite, higher probability of severe COVID-19 disease for Sex = Female, however the small value is interpreted as not significant. A score of 0 means probabilities are equal. Confidence Intervals are calculated for regression based estimates of mediation analysis metrics (NDE, TE, NIE). The statistical notion of Total Variation is calculated using Eq. (1) and Confounding Bias is derived from the Total Effect, Natural Direct and Indirect effects ($CB = TE - (NDE + NIE)$).

predictive model to this confounder bias, would predict women to be more protected from the severe COVID-19 disease forms than they really are. The Mediation Analysis on our synthetic data shows that men are expected to be more vulnerable than women 0.05 points more than they really are. Note that in reality the confounding bias can be much higher If the prediction is used to, for example, allocate limited resources in the ICU, women would be discriminated by being systematically denied priority treatment proportionate to the confounding effect.

Scenario 2: *Negative impact due to not accounting for Lifestyle mediator variables*

In this case the association between *Sex* and *COVID-19* cases is created by the mediating variable *Lifestyle* if it is not included in the data. *Lifestyle* choices such as smoking or drinking have a valid causal effect on severeness of the disease and thus, assigning a priority treatment to smoking individuals is adequate. However, if the prediction is based on *Sex* only, without observing individual patients' lifestyle habits, the women that are smokers would be wrongly classified as more resilient than they really are. As a consequence, they would be denied a part of necessary medical attention proportionate to the lifestyle Path-Specific Effect, i.e., 0.148 points higher estimated probability of severe outcome for men than for women (Table 1).

Scenario 3: *Using Biological mediators for sex-related COVID-19 severity prediction*

Considering the biological sex-specific variables such as hormones, adaptive immune systems and other variables, it is relatively safe to assume that their effect on the outcome of the disease can be predicted from the *Sex* variable. This allows for a unique situation, where using sensitive attributes is both allowed and necessary to ensure fair and accurate predictions. For example, insisting on identical treatment for men and women could result in disparate impact on health and mortality outcomes for men, proportionate to the *BioVar* Path-Specific Effect: it results in 0.521 percentage points higher probability of severe outcome for men (Table 4). Nevertheless, a careful causal path analysis is required to distinguish biological sex-related attributes from gender-related mediators or confounders that can bias the result and create unwanted discrimination at the individual or population level. In addition, individual fluctuations in biological markers can also supposedly affect constant sex-severity relationships through biological variables.

7. Discussion

The observed larger amount of hospitalized males in comparison with women, can be explained with several mediating and confounding variables. For instance, men lifestyle is different from women's, which can be a reason that men are more affected by COVID-19 infection. Men are more inclined towards drinking and smoking, which can evolve into lung infection which in turn, can formulate a larger chance of COVID-19 infection.

Social and cultural differences are additionally affecting the COVID-19 pandemic. In this line, another potential factor is the tendency of females to comply more with regulation, protecting themselves more and wearing masks more. Women are typically in charge of ensuring health for the whole family as part of their traditional reproductive work. Their greater compliance with COVID-19 recommendations is a reflection of long-established gender social roles, which has also involved an increased burden for

them during the pandemic (Power, 2020). However this behaviour does not impact the further outcome of the disease once in the hospital. Furthermore, this gender-related feature is not constant across individuals and populations.

Since the latest ML models such as deep networks do not correct against, but rather replicate existing biases of the researchers who train them, the data they are fed with, the circumstances of their testing, etc., we hope more effort is initially put into both performing representative data collection and causal data analyses. Likewise these checks need to be present when developing methods able to programmatically verify, flag, and reduce data and model biases. Stating the verified tests and/or including our recommended potential mediators and confounders will minimally set the state of affairs on the table, and therefore, highlight and make legal processes stand up for process automation. As a positive side effect, AI-based accountability will be more easily gained and traced. Fair data analysis is only the first step towards human-centric societies endowed with responsible AI systems that serve citizens and governments make use of data-informed policies more efficiently.

8. Conclusions and future work

The different impact of male and female sex on COVID-19 severity implies the existence of diversity in the different biology (dependence on the biological immune system), health status, mortality rates, lifestyle (smoking, drinking), responsibilities, and others. Our analysis suggests a possible association between smoking and a higher amount of COVID-19 deaths, as examples of a larger set of hypotheses to be studied. A responsible attitude from females towards COVID-19 prevention is another major element considered. However, to develop a better understanding of the true biological differences in disease propagation and adverse outcomes, more research is required in hormonal, inflammatory, immunologic, and phenotypical dimensions in severe COVID-19 disease.

We proposed a framework for including mediator and confounder variables identified in the literature into fair and accurate prediction models.

We use a toy linear model (Eq. (10)) to illustrate both conceptually and numerically the impact of failing to do so on the fairness of the disease severity predictions, especially in relation to the specific scenario where the prediction would be translated into a real life decision. A real world relationship between the variables should be estimated from the representative data that includes relevant mediators and confounders. We envision this task for the future work.

In this article we treated the sex and gender variables, however the consequences of disparate impact can apply to any protected variable, and they can occur in any other analysis where data collection may suffer from confounders or inappropriately interpreted mediators.

We hope this is the beginning of more methodically structured studies that consider fairness and bias from the very first data collection to the last analysis and so, lead to more fine grained causal analysis and predictions that can avoid exposing minorities at risk.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Songul Tolan and Golnoosh Farnadi for their helpful feedback. We also acknowledge Carlos Castillo for early feedback at the conceptualization phase of the paper. Finally, we thank COMETE team of Inria Saclay-Ile-de-France, Catuscia Palamidessi, Karima Makhlouf and Jasmine Nettiksimmons for useful feedback discussion and support. The work of Rūta Binkytė and Sami Zhioua is funded by the ERC grant Hypatia (https://project.inria.fr/hypatia/) under the European Union's Horizon 2020 research and innovation programme. Grant agreement № 835294. N. Díaz-Rodríguez is supported by the Spanish Government Juan de la Cierva Incorporación contract (IJC2019-039152-I), the Google Research Scholar Grant and Marie Skłodowska-Curie Actions (MSCA) Postdoctoral Fellowship with agreement ID: 101059332. Funding for open access charge: Universidad de Granada / CBUA.

References

Ahmed, S. B., & Dumanski, S. M. (2020). Sex, gender and COVID-19: a call to action. *Canadian Journal of Public Health*, *111*(6), 980–983.

Bertsimas, D., Boussioux, L., Wright, R. C., Delarue, A., Digalakis, V., Jr., Jacquillat, A., et al. (2020). From predictions to prescriptions: A data-driven response to COVID-19. arXiv preprint arXiv:2006.16509.

Besserve, M., Buchholz, S., & Schölkopf, B. (2021). Assaying large-scale testing models to interpret COVID-19 case numbers.

Binkytė-Sadauskienė, R., Makhlouf, K., Pinzón, C., Zhioua, S., & Palamidessi, C. (2022). Causal discovery for fairness. arXiv preprint arXiv:2206.06685.

Bwire, G. M. (2020). Coronavirus: Why men are more vulnerable to COVID-19 than women? *Sn Comprehensive Clinical Medicine*, 1.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 33* (pp. 7801–7808).

Chiarella, S. E., Pabelick, C., & Prakash, Y. (2021). Sex differences in the coronavirus disease 2019. In *Sex-based differences in Lung physiology* (pp. 471–490). Springer.

Dana, P. M., Sadoughi, F., Hallajzadeh, J., Asemi, Z., Mansournia, M. A., Yousefi, B., et al. (2020). An insight into the sex differences in COVID-19 patients: what are the possible causes? *Prehospital and Disaster Medicine*, *35*(4), 438–441.

De La Vega, R., Barquín, R. R., Boros, S., & Szabo, A. (2020). Could attitudes toward COVID-19 in Spain render men more vulnerable than women? PsyArXiv.

Ebadi, M., & Montano-Loza, A. J. (2020). Perspective: improving vitamin D status in the management of COVID-19. *European Journal of Clinical Nutrition*, *74*(6), 856–859.

Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R., & Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of Sex Differences*, *11*(1), 1–13.
Geiger, D., & Pearl, J. (1990). On the logic of causal models. In *Machine intelligence and pattern recognition. Vol. 9* (pp. 3–14). Elsevier.
Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., & Sebag, M. (2018). Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning* (pp. 39–80). Springer.
Grzelak, L., Velay, A., Madec, Y., Gallais, F., Staropoli, I., Schmidt-Mutter, C., et al. (2020). *Sex differences in the decline of neutralizing antibodies to SARS-CoV-2*. Cold Spring Harbor Laboratory Press, MedRxiv.
Head, J. R., Andrejko, K., Cheng, Q., Collender, P. A., Phillips, S., Boser, A., et al. (2020). *The effect of school closures and reopening strategies on COVID-19 infection dynamics in the San Francisco Bay Area: a cross-sectional survey and modeling analysis*. Cold Spring Harbor Laboratory Press, MedRxiv.
Jain, A., Chaurasia, R., Sengar, N. S., Singh, M., Mahor, S., & Narain, S. (2020). Analysis of vitamin d level among asymptomatic and critically ill COVID-19 patients and its correlation with inflammatory markers. *Scientific Reports*, *10*(1), 1–8.
Kabir, H., Maple, M., Usher, K., & Islam, M. S. (2019). Health vulnerabilities of readymade garment (RMG) workers: a systematic review. *BMC Public Health*, *19*(1), 1–20.
Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems* (pp. 656–666).
Klein, S. L., Dhakal, S., Ursin, R. L., Deshpande, S., Sandberg, K., & Mauvais-Jarvis, F. (2020). Biological sex impacts COVID-19 outcomes. *PLoS Pathogens*, *16*(6), Article e1008570.
Klein, S. L., Jedlicka, A., & Pekosz, A. (2010). The Xs and Y of immune responses to viral vaccines. *The Lancet Infectious Diseases*, *10*(5), 338–349.
Kopel, J., Perisetti, A., Roghani, A., Aziz, M., Gajendran, M., & Goyal, H. (2020). Racial and gender-based differences in COVID-19. *Frontiers in Public Health*, *8*, 418.
Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems* (pp. 4066–4076). USA.
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions. arXiv preprint arXiv:2010.09553.
Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420).
Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.
Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459.
Pearl, J., & Dechter, R. (2013). Identifying independencies in causal graphs with feedback. arXiv preprint arXiv:1302.3595.
Peckham, H., de Gruijter, N. M., Raine, C., Radziszewska, A., Ciurtin, C., Wedderburn, L. R., et al. (2020). Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nature Communications*, *11*(1), 1–10.
Pourali, F., Afshari, M., Alizadeh-Navaei, R., Javidnia, J., Moosazadeh, M., & Hessami, A. (2020). Relationship between blood group and risk of infection and death in COVID-19: a live meta-analysis. *New Microbes and New Infections*, *37*, Article 100743.
Power, K. (2020). The COVID-19 pandemic has increased the care burden of women and families. *Sustainability: Science, Practice and Policy*, *16*(1), 67–73.
Sharma, G., Volgman, A. S., & Michos, E. D. (2020). Sex differences in mortality from COVID-19 pandemic: are men vulnerable and women protected? *Case Reports*, *2*(9), 1407–1410.
Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(10).
Silpasuwan, P., Prayomyong, S., Sujitrat, D., & Suwan-Ampai, P. (2016). Cotton dust exposure and resulting respiratory disorders among home-based garment workers. *Workplace Health & Safety*, *64*(3), 95–102.
Smith, T. (2020). A supercomputer analyzed COVID-19—and an interesting new theory has emerged. Elemental https://elemental.medium.com/a-supercomputer-analyzed-covid-19-and-an-interesting-new-theory-has-emerged-31cb8eba9d63.
Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, *9*(1), 62–72.
Starkopf, L., Andersen, M., Gerds, T., Torp-Pedersen, C., & Lange, T. (2017). *Comparison of five software solutions to mediation analysis*. Copenhagen, Denmark: University of Copenhagen.
Tadiri, C. P., Gisinger, T., Kautzy-Willer, A., Kublickiene, K., Herrero, M. T., Raparelli, V., et al. 0000. The influence of sex and gender domains on COVID-19 cases and mortality.
Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis*. UCLA Statistics/American Statistical Association.
Traish, A. M., & Morgentaler, A. (2021). What's testosterone got to do with it? A critical assessment of the contribution of testosterone to gender disparities in COVID-19 infections and deaths. *Androgens: Clinical Research and Therapeutics*, *2*(1), 18–35.
von Kügelgen, J., Gresele, L., & Schölkopf, B. (2020). Simpson's paradox in COVID-19 case fatality rates: a mediation analysis of age-related causal effects. arXiv preprint arXiv:2005.07180.
Wenham, C., Smith, J., & Morgan, R. (2020). COVID-19: the gendered impacts of the outbreak. *The Lancet*, *395*(10227), 846–848.
Wu, Y., Zhang, L., Wu, X., & Tong, H. (2019). Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in neural information processing systems* (pp. 3404–3414).
Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).
Zeberg, H., & Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, *587*(7835), 610–612.
Zhang, J., & Bareinboim, E. (2018). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1*.
Zhou, X., & Yamamoto, T. (2020). Tracing causal paths from experimental and observational data. SocArXiv.
Zietz, M., & Tatonetti, N. P. (2020). *Testing the association between blood type and COVID-19 infection, intubation, and death*. Cold Spring Harbor Laboratory Preprints, MedRxiv.
Zietz, M., Zucker, J., & Tatonetti, N. P. (2020). Associations between blood type and COVID-19 infection, intubation, and death. *Nature Communications*, *11*(1), 1–6.