



Extending the Range of COVID-19 Risk Factors in a Bayesian Network Model for Personalised Risk Assessment

Prodhan G* and Fenton N

School of Electronic Engineering and Computer Science, University of London, UK

*Corresponding author: Georgina Prodhan, School of Electronic Engineering and Computer Science Queen Mary, University of London, 58 Dorville Road, London SE12 8EB, UK, Tel: +447833435262; Email: g.m.prodhan@se19.qmul.ac.uk

Conceptual Paper

Volume 4 Issue 6

Received Date: October 23, 2020

Published Date: November 20, 2020

DOI: 10.23880/eij-16000170

Abstract

A need is emerging for individuals to gauge their own risks of coronavirus infection as it becomes apparent that contact tracing to contain the spread of the virus is not working in many societies. This paper presents an extension of an existing Bayesian network model for an application in which people can add their own personal risk factors to calculate their probability of exposure to the virus and likely severity if they do catch the illness. The data need not be shared with any central authority. In this way, people can become more aware of their individual risks and adjust their behavior accordingly, as many countries prepare for a second wave of infections or a prolonged pandemic. This has the advantage not only of preserving privacy but also of containing the virus more effectively by allowing users to act without the time lag of waiting to be informed that a contact has been tested and confirmed COVID-19 positive. Through a nuanced assessment of individual risk, it could also release many people from isolation who are judged highly vulnerable using cruder measures, helping to boost economic activity and decrease social isolation without unduly increasing transmission risk. Although much has been written and reported about single risk factors, little has been done to bring these factors together in a user-friendly way to give an overall risk rating. The causal probabilistic model presented here shows the power of Bayesian networks to represent the interplay of multiple, dependent variables and to predict outcomes. The network, designed for use in the UK, is built using detailed data from government and health authorities and the latest research, and is capable of dynamic updates as new information becomes available. The focus of the paper is on the extended set of risk factors.

Keywords: Coronavirus; COVID-19; Contact Tracing; App; Risk; Privacy; Second Wave

Abbreviations: COPD: Chronic Obstructive Pulmonary Disease; PPE: Personal Protective Equipment; ICUs: Intensive Care Units; BMI: Body Mass Index.

Introduction

Like many countries, Britain is bracing for a second wave of coronavirus. Local lockdowns are being reimposed in areas of England, which suffered the highest levels of excess mortality in Europe [1], even as business are encouraged to stay open, individuals urged to go back to work in offices,

and schools go back for the new academic year. Mitigating measures such as the compulsory wearing of face masks in confined public spaces have recently been imposed. But a digital contact-tracking system that was supposed to be the basis for a lifting of restrictions on movement and social activity has yet to prove it effective. The UK government has just launched a revamped version of its coronavirus app. It has also been reported to be considering a feature that allows people to use personal information to calculate their own risk score [2]. The model presented here could be the basis for just such an app. The location of new outbreaks could

be identified simply by collecting data on users' location (via GPS, not the more invasive Bluetooth) and likelihood of infection.

This approach to containing the virus relies largely on individuals taking responsibility for their own health and that of those around them. This may be thought to be a weakness. But in fact, in all but the most authoritarian regimes, a sense of social responsibility is essential if rules on social distancing, face masks, hand-washing and other measures are to be effective. Globally, experience of past pandemics including the 2009 swine flu, SARS, the 2014-2016 Ebola outbreak and various outbreaks of bird flu have shown that people's behaviour-based on risk perception - can have a profound influence on the spread of infectious diseases [3].

There is evidence that the model presented here could be effective in Britain (although it is, of course, adaptable to any country in the world). A study in 10 countries that was published in May Dryhurst S, et al. [4] found that the UK had the highest public risk perception of COVID-19 in the group. The COVID Symptom Study, an epidemiological research app that invites users to report any symptoms and share details of their health lifestyle, has more than 4 million contributors, mostly in the UK [5]. And, according to a poll of 2,254 UK residents carried out in May by King's College and Ipsos MORI [6], the majority has been going beyond compliance with government advice, for example by staying at home for long periods. The poll also found that most people would obey a contact-tracing app's recommendation to self-isolate - showing the potential utility of an app that relies on voluntary participation. However, the majority did not trust the government to keep their data safe, and almost half were Sceptical about the ability of such an app to limit the spread of coronavirus.

This paper is part of a set of studies conducted by a multidisciplinary team of researchers who argue that contact tracing alone is unlikely to contain a high-prevalence, contagious disease such as COVID-19. It builds on the work most recently of Fenton N [7] which describes details of a Bayesian model to compute the probability that an individual has COVID-19 - whether symptomatic or asymptomatic - or is likely to catch it. The model presented here expands substantially on that web of relevant risk factors, bringing in ethnicity, religion, occupation and housing conditions, and refining other factors such as age, obesity and underlying medical conditions. It also incorporates concurrent work on an extended set of symptoms described in Butcher R, et al. [8].

Results of running the model are consistent with what is known about the prevalence and severity of COVID-19, and

the vulnerability of different groups. This demonstrates its ability to predict risks for individuals, even if relatively little is known about them. For example:

- Entering the observation that someone is Black or non-Chinese Asian results in a raised probability of having severe COVID-19 in the 55-64 age groups and an equal probability in every other age group. This is consistent with observed data [9]. The opposite is true, however, when comparing ethnic groups without conditioning on age- an interesting paradox that will be discussed below.
- Entering the observation that someone is aged under 16 reduces the probability of having severe COVID-19 to virtually zero but raises the probability of having the disease asymptotically. This is consistent with observed data and an important factor to watch as schools reopen [10]. Adding the observation "student/pupil" raises the probability of eventual COVID-19.
- The probability of infection takes precedence over background risks - living in an overcrowded household is more dangerous than being obese. This is logical, since nobody can have COVID-19 without becoming infected, no matter what their vulnerabilities. It shows the power of mitigating measures, and cautions against overemphasizing personal risks such as obesity or smoking.

Context and Related Work

Countries around the world have introduced contact-tracing apps in attempts to contain a second wave of coronavirus as lockdowns are eased. Contact tracing using mobile technology has become a particular focus during the current epidemic because COVID-19 presents the challenge that the majority of transmission is believed to occur before symptoms of the disease show. This differentiates it from ebola, smallpox or the 2002 SARS-CoV-1, and makes it vital to reduce the time from symptoms appearing in a person to isolation of that person's contacts. Recent studies including Ferretti L, et al. [11], Kretzschmar ME, et al. [12] suggest that the time from confirmation of a COVID-19 case to quarantining of contacts needs to be less than three days to bring the rate of reproduction below 1, thus containing the virus's spread. Such a time frame is generally beyond the capabilities of manual tracers, especially for an epidemic of the current scale. The study, which examined early epidemic data from China, Singapore and the Diamond Princess Cruise ship, concluded that near-instantaneous contact tracing via a mobile phone app, in conjunction with other measures, could contain the virus.

While adoption of mobile contact-tracing apps has been relatively high and has been credited with helping to flatten the curve in some East Asian countries, the same conditions

do not apply in much of the West [13]. In China, using a traffic-light health-code system embedded in the extremely popular WeChat and Alipay platforms is practically mandatory for freedom of movement. In South Korea, which swiftly contained the first outbreak (although it is now battling a second), authority had already been given to disease-prevention authorities to override some privacy laws during the 2015 MERS outbreak [14]. Crucially, the country also quickly developed large-scale testing infrastructure. In Singapore, the first country to deploy a national coronavirus-tracing app, a wearable token is being rolled out to address weaknesses in the existing mobile-phone technology – and has sparked a rare backlash against the government, with accusations it is on the way to becoming a surveillance state [15].

Among Western countries, adoption of government-endorsed mobile contact-tracing apps has been relatively high in Australia, Germany and Italy, with rates of around 22%, 14% and 7% respectively by mid-July [16]. But none of these countries can show or is willing to claim the effectiveness of the app in limiting the spread of the virus. And even the countries with the highest adoption are still well below the threshold of 56% of the population, or 80% of smartphone users, that experts consider the minimum for the technology to have a chance of being effective [17]. Privacy concerns are largely to blame, along with technical glitches and discomfort with digital technology among some groups, especially older people.

Some privacy-preserving contact-tracing apps are already on the market or in development. A few stand out:

- **Private Kit: Safe Paths** developed at Massachusetts Institute of Technology. This is an open-source technology that provides individual users information on their interaction with COVID-19, and allows them to share their location trails with health officials if they test positive for the disease [18]. The app is available for iOS and android.
- **COVID-19 Watch** from Stanford University. This is an early-warning system using Bluetooth, which allows an infected person to send an anonymous alert to others they may have infected, if they are also users of the app [19]. It is currently in testing.
- **CoEpi.** This is a community-based epidemiological tool that uses Bluetooth technology and is based on voluntary symptom-sharing, including before confirmed test results [20]. It is currently in development.

All of these apps and proposed apps address privacy concerns, and CoEpi also potentially collapses the time from

infection to isolation of a contact by promoting symptom-sharing before test results. They all deal exclusively with infection risks, however, and none incorporates background risk to give a personalised risk assessment in the way that the model presented here does.

I searched the preprint servers medRxiv, bioRxiv and arXiv as well as Google Scholar and LitCovid for Bayesian approaches to risk assessment and contact tracing. I found several interesting papers on transmission dynamics and mitigation measures such as physical distancing, face masks and eye coverings. An early study of Chinese data formulated an equation to determine personal risk based on age, sex and comorbidities [21]. It used a naïve Bayesian approach, assuming that all the risk variables were independent of one another. An interesting French study modelled individual risk as a potential decision-making aid to relaxing lockdown restrictions on low-risk populations [22]. It used standard statistical and machine-learning methods such as logistic regression and random forest. Another, Italian, paper aimed to develop a personal risk score for infection [23]. However, it only measured association, not causality, which was evident in its finding that underlying medical conditions could influence the risk of infection. In causal Bayesian terms, this would be a category mistake as the most vulnerable person in the world could not become infected with coronavirus if they did not come into contact with it. I found nothing that used a Bayesian analysis for personal risk prediction and containment, apart from the work on which this project is based Fenton N, et al. [24].

Methodology

The model presented is a Bayesian Network, a causal probabilistic model that predicts outcomes—in this case, current and projected COVID-19 status and severity of disease using a combination of actual observations and expert calculations. It is built on AgenaRisk Bayesian network software. A simplified schema of the network is shown in Figure 1, and a more detailed version can be found in the Appendix. The model, named “Combined Model for Contact Tracing” can be accessed using this link <http://www.eecs.qmul.ac.uk/~norman/Models/> and can be run using trial software from agenarisk.com.

The network is initially populated with prior probabilities for the general population—e.g., a person in the UK has a 49.4% chance of being male, a 23.4% chance of being obese and a 16% chance of being a smoker. Together with other factors including infection opportunities, and immunity possibility, these resulted in a 0.2% chance of eventually having COVID-19, at the time of writing.

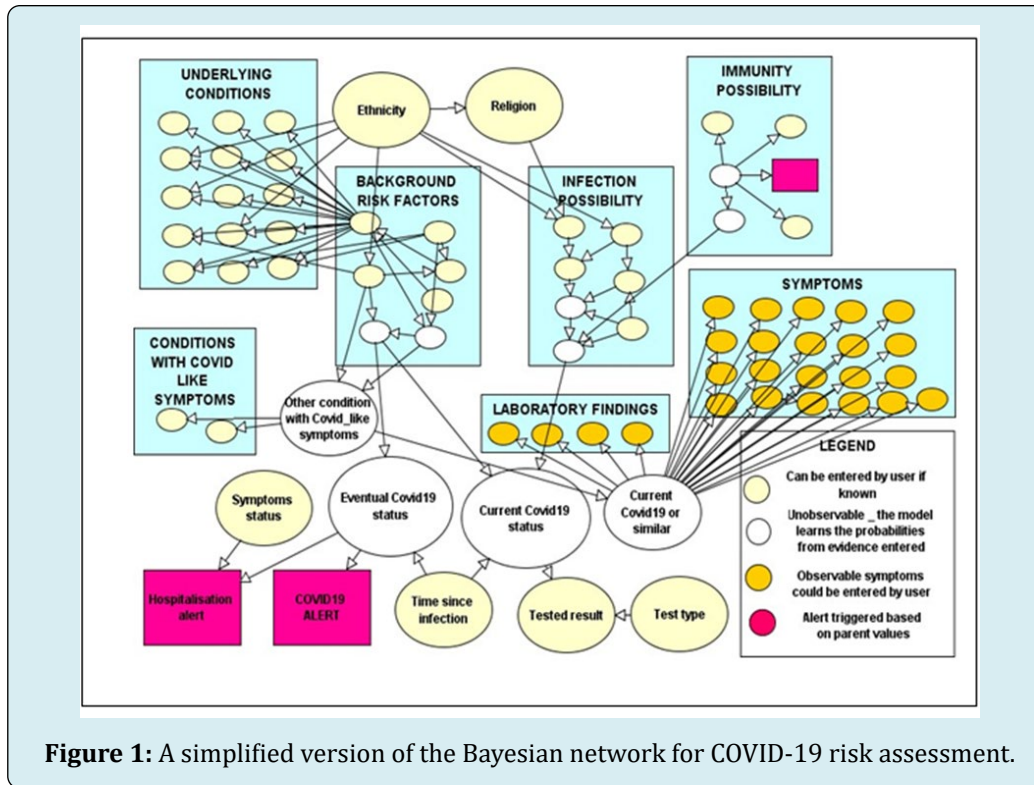


Figure 1: A simplified version of the Bayesian network for COVID-19 risk assessment.

Bayesian networks are known to be able to solve problems that traditional statistics cannot, notably cases of confounders and colliders-in which associated variables are causally related only through a common cause or a common effect. A Bayesian approach is well-suited for the current situation, in which it is crucial to understand the relationships of multiple dependent variables, and where observational data have to be gleaned opportunistically as controlled experimentation is impossible.

Prime examples of confounding variables surround the issue of ethnicity. Several large studies including Williamson EJ, et al. [25], Office for National Statistics [26] found a significantly higher risk of death from COVID-19 among Black or South Asian people than among White people. But by looking only at association or correlation, their methods were by definition unable to discover the mechanisms by which this occurred. Even after it said it had adjusted for age, region, population density, area deprivation, household composition, socio-economic position, education, household tenure, multigenerational households and occupation, the ONS still found that Black males had twice the risk of White males, which it found to be “unexplained”.

In the Bayesian network presented here, it can be clearly seen that the relationship between ethnicity and COVID-19 death is in fact confounded by multiple variables: Black and South Asian people are more likely to be frontline healthcare

workers, live in overcrowded households or have sickle cell disease, for example, than White people. The healthcare worker case is illustrated in Figure 2. All of these are factors that increase the risk of catching COVID-19 or may lead to a more severe form of the disease. They confound the relationship between ethnicity and COVID-19 outcomes.

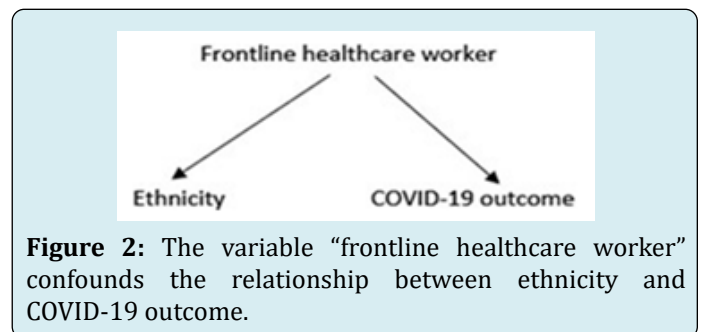


Figure 2: The variable “frontline healthcare worker” confounds the relationship between ethnicity and COVID-19 outcome.

Potential collider variables are also abundant in the current COVID-19 situation because of the nature of the studies that have been carried out. Observational data cannot be randomly sampled and controlled experiments cannot be done so almost all the data available come from hospital populations, individuals who have been tested or people voluntarily taking part in studies. These are not representative of the population: hospitals patients are already frail, tested individuals are already exposed or symptomatic and voluntary participants may tend to be health-conscious,

well-educated, tech-savvy or hypochondriac. As such, any study conducted on one of these populations is already conditioned on a collider variable such as “frail”, “vulnerable” (e.g. frontline healthcare worker) or “tech-savvy” (i.e. likely to be younger). A UK government community infection pilot study [27], which conducts regular tests on a large number of randomly selected participants, is beginning to redress this situation to some extent.

Observing associations between other factors in such pre-selected conditions can result in misleading conclusions. Many of the underlying medical conditions associated with more severe COVID-19 may bear no causal relation to severity of COVID-19 but simply have been observed in hospitalized patients with severe COVID-19 because those conditions make them more likely to be hospitalized in the first place. For example, it is not completely clear how heart disease may influence COVID-19 outcome, but by conditioning studies on the collider variable “hospitalized” we are confining our observations to a population already more likely to have heart disease or a whole range of other co-morbidities, and ignoring people who have those conditions combined with mild COVID-19 but remain out of hospital. In this way, an exaggerated or even a false association can appear between the variables “heart disease” and “severe COVID-19”, as seen in Figure 3. The U.S. Centers for Disease Control and Prevention has noted that only 6% of the COVID-19 deaths so far have mentioned COVID-19 as the sole cause of death [28].

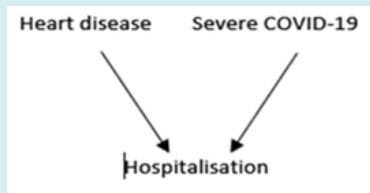


Figure 3: Heart disease and severe COVID-19 collide on the variable “hospitalization”.

Underling medical condition	High						Medium						Low						
	Ethnicity	White	Mixed/...	Chinese	Other...	Black	Other	White	Mixed/...	Chinese	Other	Black	Other	White	Mixed/...	Chinese	Other...	Black	Other
FALSE	0.834	0.834	0.847	0.674	0.725	0.834	0.834	0.834	0.847	0.674	0.73	0.834	1	1	1	1	1	1	1
TRUE	0.166	0.166	0.153	0.326	0.275	0.166	0.166	0.166	0.153	0.326	0.28	0.166	0	0	0	0	0	0	0

Table 2: Node probability table for “diabetes”.

The structure of the network, inherited from Fenton N, et al. [24], consists of four main areas that feed into the calculation of probability of COVID-19: background risk factors, infection possibility, immunity possibility, and symptoms. I developed the areas of background risk factors

In the current version of the model, I have included the underlying medical conditions that best expert advice considers linked with higher risk of severe COVID-19 because of a lack of further evidence to causally explain them or explain them away. But the list has changed many times during the making of the model and may change again.

It is worth noting that even though male sex is widely observed and accepted as being associated with higher risk, this has not yet been convincingly explained in a causal way and may yet turn out to be a relationship that is confounded by other factors. It has been suggested that the hormone oestrogen (which men have in smaller amounts) is a protective factor. A new study argues that females may mount a more robust immune response than males, even in older age [29]. But research is still at an early stage. Again, a Bayesian model can easily adapt to new knowledge, for example by creating a new node “oestrogen level” and drawing edges connecting this to both males and females, with different prior probabilities.

Model Architecture

The Bayesian network is an acyclic directed graph that consists of probability nodes connected by edges representing conditional dependencies. Inside each probability node is a probability table. Some of these are simple, e.g. for sex or for age, which are not conditioned on any other factors-the nodes have no “parents”-as seen in Table 1. In other cases, they are more complex, e.g. the diabetes node probability table is conditioned on underlying medical condition and ethnicity-as seen in Table 2.

Male	Female
0.4594	0.506

Table 1: Node probability table for “sex”.

and infection possibility. My colleague Rachel Butcher developed the symptoms area.

Entering an observation in any of the nodes will change the probabilities of all the variables that are connected

to it, whether in the direction of influence or by backward inference. For example, entering an observation that a person is aged 75 or older will raise the probability of currently having severe COVID-19 from 0.01% to 0.02% and of eventual severe COVID-19 from 0.03% to 0.06%. But it will also raise the probability of being White from 84.8% to 95.9% (because White people live longer), it will raise the probability of being Christian from 58.7% to 62.6% (because more White people are Christian) and it will decrease the probability of being morbidly obese from 2.8% to 1.5% (because some of the morbidly obese will have died by that age, and elderly people tend to eat less). As many facts as are known can be entered - the more observations, the more accurate the prediction. A few of these nodes and how their probability tables were arrived at will be discussed below. Details of all the node probability tables that I developed are in the Appendix, along with their data sources.

Background Risk Factors

At the time of writing, there was broad consensus that age, sex, obesity and certain underlying medical conditions were important risk factors for developing more severe forms or dying of COVID-19 [31-35]. The main difficulty arose in how to represent the interplay of the various factors, given that age, sex and obesity were themselves interrelated and in addition could all influence the probability of underlying medical conditions. It would be impossible to give accurate probabilities for each combination of these (e.g. probability of having an underlying medical condition for an overweight woman aged 35-54, and then the probability of an underlying medical condition for that same woman if she were normal/underweight, obese or morbidly obese, and so on). In fact, for seven age categories, two sex categories and four obesity categories there would be 56 combinations of factors that could each lead to a high, medium or low probability of having an underlying medical condition. Clearly, it would be over-engineering the model to give a value to each of these.

Instead, I drew links between nodes where I did have accurate figures. For example, there were reliable data for obesity by age group and sex so I made age and sex parent nodes of obesity. I made particular underlying medical conditions child nodes of the general underlying medical conditions node, with age, obesity and sex as parent nodes of those individual conditions where their influence was known to be strong.

To capture the complexity of the influence of ethnicity on COVID-19 risk, the node "ethnicity" was placed outside the background factors section of the model as a parent of other factors such as age, underlying medical condition and overcrowded housing (an infection factor).

Despite a flurry of excitement around a French study released in June that proposed smoking may have a protective effect against developing COVID-19 symptoms [36], this was soon dismissed by expert consensus, and the direct effects of smoking on COVID-19, if any, are still unknown [37]. Smoking was therefore included in the model as a parent node of "underlying medical condition" and also of chronic obstructive pulmonary disease (COPD).

Infection Possibility

Occupation was the single most important factor in determining the probability of exposure to the virus, with frontline healthcare workers clearly the most exposed. South Asian and Black people are overrepresented in the UK's National Health Service and therefore especially in early COVID-19 deaths, when personal protective equipment (PPE) was scarce.

Living in multi-generational households is also a risk factor, especially in cases where a middle generation goes out to work and lives with elderly relatives and with children who may socialize more outside the home. This is more common in South Asian communities but is also becoming generally more popular as elderly people move in with their children, while adult offspring move back or stay in their parental home for economic reasons. I drew a link from ethnicity to overcrowded household, as data exist for overcrowding by ethnicity. I also drew a link from religion to overcrowded household, mainly to account for the large multi-generational households of Britain's strictly Orthodox Haredi population, which was not captured by ethnicity [38].

Selected Probability Nodes

Ethnicity

The modelling of ethnicity was one of the most complex parts of building the network. A whirlwind of media and statistical reports was published in Britain and the United States regarding the higher vulnerability of non-White ethnic groups to severe COVID-19 symptoms, hospitalization and death in the wake of the killing of George Floyd and subsequent surge in support for the Black Lives Matter movement that occurred while I was working on the model. No substantial evidence has as yet emerged of genetic factors that would explain a racial bias in the spread of the disease. Genomics England and the University of Edinburgh have launched a nationwide study to understand how a person's genetic makeup could influence how they react to the virus. [39]. This is still in progress.

The media and political storm did cause a lot of data to be published analyzing COVID-19 statistics by ethnicity, which

was very useful in the building of this model. Therefore, I was able to represent the relationship between ethnicity and severity of COVID-19 risks with some accuracy through a greater tendency of ethnic minorities in the UK to live in

overcrowded housing, suffer from certain health conditions such as diabetes, and work on the healthcare frontline. I have reproduced below a visualization of the complexity of the situation by Pareek M, et al. [30] in Figure 4.

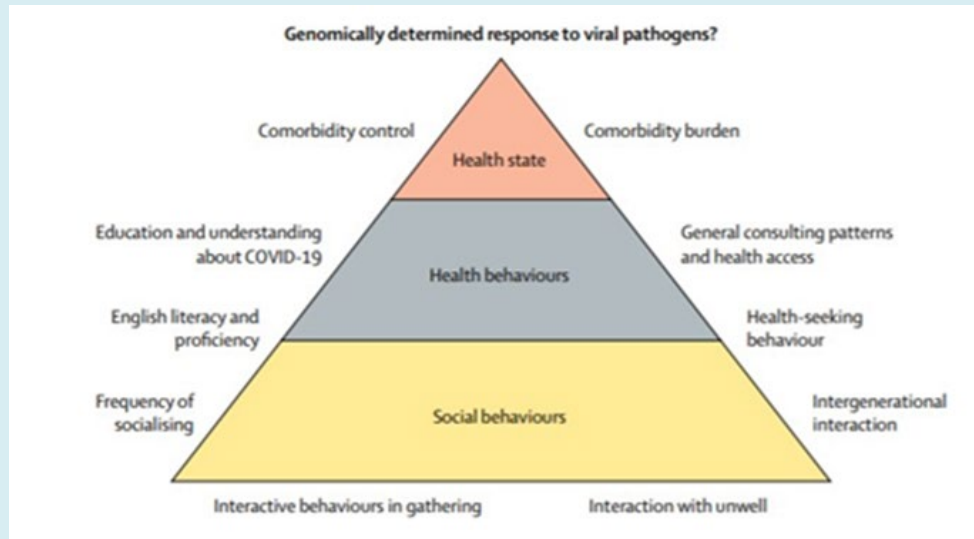


Figure 4: The potential interaction of ethnicity-related factors on SARS-CoV-2 infection likelihood and COVID-19 outcomes [30].

Even with only a few of these important factors captured, the model is capable of assessing that a White, female nurse of normal weight living in shared accommodation is at higher risk than a Black, obese male office clerk working from home, where he lives with his nuclear family—even though being Black and obese may overall be likely to make for higher risk

in the absence of other knowledge. A fragment of the model showing the edges connecting ethnicity to other nodes is shown in Figure 5. Note that some of these factors – short lifespan, overcrowded household, risky occupation—allow for interventions to make matters more equitable. Ethnicity is of course not a variable that can be changed.

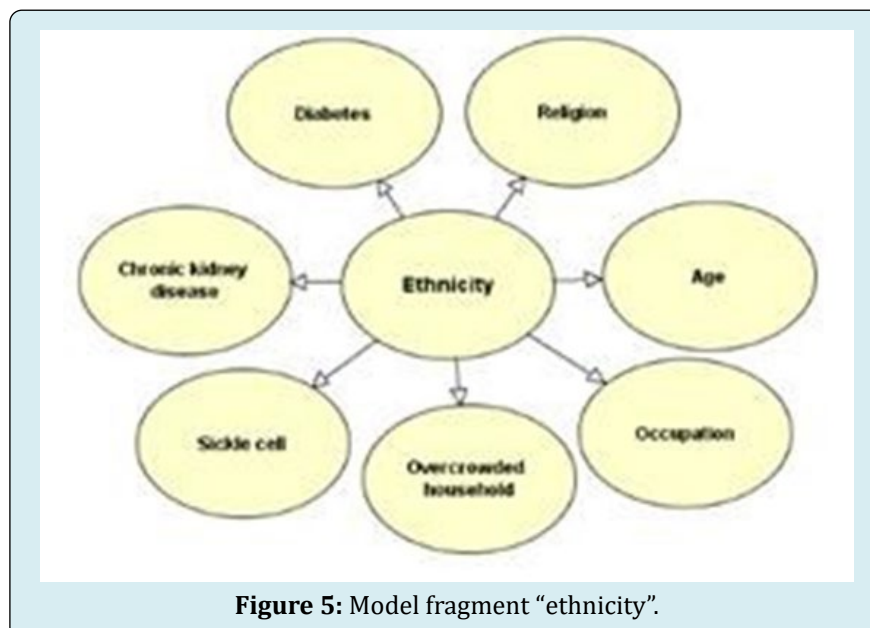


Figure 5: Model fragment "ethnicity".

The connection between age and ethnicity is an important one. It explains the paradoxical situation that Black and non-Chinese Asian people have a lower risk of developing severe COVID-19 when considered as whole populations – in apparent contradiction of the news headlines – but they do have an equal or increased risk when considered age group by age group. This is an example of Simpson's Paradox, in which a characteristic of a larger group can disappear or be reversed when considering subgroups one by one. The situation is summarized in Table 3.

Age	White	Asian*	Black
Unobserved	0.03%	0.02%	0.02%
75+	0.06%	0.06%	0.06%
65-74	0.05%	0.05%	0.05%
55-64	0.03%	0.04%	0.04%
35-54	0.02%	0.02%	0.02%
25-34	0.02%	0.02%	0.02%
16-24	0.01%	0.01%	0.01%
<16	0.01%	0.01%	0.01%

*excluding Chinese

Table 3: Probability of Eventual Severe Covid-19 Illness by Ethnicity and Age.

The reason for this is simply that White people, who make up the majority of the UK's population, tend to live longer than other ethnic groups. Age is also the single most important determinant of COVID-19 severity. Taken overall, the UK's Black and non-Chinese Asian populations are relatively young, with only a small proportion living long enough to fall into higher-risk age categories. Considering the sub-categories of age group rather than the whole population is therefore the right way to analyse the situation, comparing like with like. This is a good validation of the model. The effect is similar but more pronounced in the United States, where infection rates are still significantly higher than in Britain [40].

Obesity

Certain underlying medical conditions are widely believed by experts to contribute to the risk of developing more severe COVID-19 symptoms or dying of the disease. Some large-cohort studies of hospital patients have analysed factors associated with COVID-19 death [41]. These are observational studies and, as such, do not establish causal relationships. Nonetheless, they represent best expert opinion at the time of writing. Many experts believe that,

broadly, underlying medical conditions weaken the immune system and thus limit the body's ability to fight off infection or illness.

One such condition is obesity, and in late July the UK government launched a campaign to fight it, arguing that people could make themselves less vulnerable to a second wave of coronavirus by losing weight. Public Health England published a long report on the topic, which synthesized a large amount of useful information but was flawed in many ways. For example, it argued:

“Evidence suggests excess weight is associated with an increased risk of the following for COVID-19: a positive test, hospitalization, advanced levels of treatment (including mechanical ventilation or admission to intensive or critical care) and death”.

It is unclear why it would be worth noting that excess weight is associated with a positive test since, as the report itself acknowledges, obesity does not increase the chance of infection. Secondly, the report highlighted the fact that 31.3% of patients critically ill in intensive care units (ICUs) were obese, with a body mass index (BMI) of 30 or over, compared with 28.9% of the population, while 7.9% were morbidly obese, with a BMI of 40 or over, compared with 2.9% of the general population. Both were adjusted for age and sex in unspecified ways. Aside from the fact that the difference in the case of simple obesity is very small, this gives a misleading impression of the risks of obesity when reading the results in the causal direction—from obesity to intensive care. In my model, entering the observation “severe” in the “current COVID status” node also raises the probabilities of being morbidly obese or obese, as seen in Table 4. (The data used differ slightly from those in the PHE report.)

	Prior probability	Observe current COVID-19 status severe
Morbidly obese	2.85%	3.47%
Obese	23.35%	27.84%
Overweight	31.05%	33.95%
Normal/underweight	42.75%	34.75%

Table 4: Probability of Obesity Given Severe Covid-19.

However, reading the data in the direction of influence by clearing the observation “severe COVID-19” and observing obesity and morbid obesity in turn has no perceptible effect on COVID-19 status measured to two decimal places, as seen in Table 5.

	Prior probability	Observe obese	Observe morbidly obese
COVID-19 status severe	0.03%	0.03%	0.03%
COVID-19 status mild	0.11%	0.11%	0.11%
COVID-19 status asymptomatic	0.07%	0.06%	0.06%
COVID-19 status none	99.80%	99.80%	99.80%

Table 5: Probability of Eventual Covid-19 Given Obesity.

This shows that backward inference is not a guide to effectiveness of intervention and, while losing weight is certainly good health advice for many people, an anti-obesity drive may be a waste of public time and money as a measure to fight the effects of any second coronavirus wave. I included obesity in the model as a parent of “underlying medical condition” as, like smoking, it is a predictor of several conditions. I did not include it as a parent of “risk factors” in its own right.

Occupation

Occupation is a key determinant of exposure to coronavirus. Frontline healthcare workers are at significantly higher risk because they are exposed to many people, and those people are far more likely than the general population to be ill with COVID-19. Other essential workers who have contact with others at work or who have to travel to their place of work on public transport are also significantly more exposed than the general population. The UK’s Office for National Statistics has analysed occupations with the highest potential exposure. In some cases, numbers are small and so this is a work in progress. Risky occupations outside healthcare included bus drivers care home workers and home carers, security guards, primary and nursery teachers, police officers, opticians and pharmacists, plumbers, vets and undertakers. I included students/pupils as an occupational category in their own right, as COVID-19 outbreaks began to appear on campuses and in schools at the start of the new academic term.

It is envisaged that users of the app would have a multiple-choice question about their occupation, which would yield a high, low or medium risk category. The actual list of risky occupations could be adjusted as more research became available.

The ONS analysed those employed in these areas by sex and ethnicity. Some occupations were skewed towards women (carers and teachers), some towards South Asians (pharmacists and opticians) and some towards white males (police officers and plumbers). Overall, there seemed no great imbalance in the chances of belonging to one of these medium-risk occupational groups so I did not consider ethnicity or sex as parents of those. I did, however, do so for

frontline health workers, using data from the NHS, which has critical mass with 1.24 million employees, making it the biggest employer in the UK, and where Black and south Asian employees are overrepresented.

Scenarios

A few fictional cases illustrate the power of this model over simple contact-tracing exposure models, and also its ability to give a risk assessment far more useful to individuals than the general government guidance.

The Unemployed White Male

Fred is a 55-year-old unemployed White man. He is obese and he smokes but he has no known medical conditions. He lives alone and has minimal interactions with people outside the home. He describes his religion as “none”. He has not knowingly had any contact with people with COVID-19 or COVID-19 symptoms. He worries about his weight and his smoking but in fact, the model predicts, his risk of eventual severe COVID-19 is lower than average and he is strongly predicted not to get COVID-19 at all, with a probability of 99.85% [42].

The Retired Female Country-Dweller

Katie is a 79-year-old, white, retired lawyer. She has a touch of Parkinson’s disease but is of a healthy weight and has never smoked. She lives in a small village with her husband and has had minimal contact with other people since the lockdown. She is Catholic and goes once a week to church, where she socially distances. Katie is worried about her vulnerability, mainly because of her age, and follows government advice to shield at home, not even venturing out for walks. The model calculates her probability of eventual severe COVID-19 if she gets it as above average at 0.07%, but also a 99.85% chance she will not become infected at all.

The Urban Asian Bus Driver

Rajiv is a 35-year-old London bus driver of Indian origin. He is a fitness enthusiast and takes care to maintain a healthy weight. Despite that, he has type 1 diabetes. He gave up smoking last year. Rajiv does not observe any religion, and he lives in a comfortable home with no overcrowding.

He takes care not to mix with other people too much outside of work, and he feels confident he will not become too ill even if he does catch COVID-19 because of his youth and his commitment to fitness. The model predicts his probability of eventually contracting severe COVID-19 at 0.04%, slightly higher than that of the general population. He is more likely than the average person to contract some degree of COVID-19, with a 99.73% chance of not contracting it at all.

The African Hospital Orderly

Victor is a 29-year-old Oxford hospital orderly, Black, of Ghanaian origin. His weight is normal, he has never smoked and has no known medical conditions. He goes regularly to an evangelical church but otherwise avoids other people at the moment, although he has to use public transport to travel to work. He lives with a flat mate who works from home. He has been in contact with patients with COVID-19 symptoms at work although he does not work in a COVID-19 ward and has not knowingly been in close contact with a confirmed COVID-19 patient. The model predicts a 0.2% chance of severe COVID-19-10 times that of the general population and only a 97.64% chance of not contracting the disease at all.

The Chinese Schoolgirl

Yun Chee is an 11-year-old Chinese schoolgirl living in London. She is of normal weight, has no underlying medical conditions, and is not religious. She had only been to school a few times in the past months until her school reopened in September. She has also started going on playdates and sleepovers with her friends. She lives in a small house with her parents and grandparents, with one bathroom and one kitchen. Yun Chee has a lower-than-average probability of 0.01% of contracting severe COVID-19, the model predicts, but almost double the average probability of eventually having it asymptotically, at 0.13%. This is a risk that should be monitored closely and may need to be adjusted according to observational data as schools remain open.

Conclusion

A Bayesian network such as the one presented here could be a powerful tool for evaluating individual risk as countries around the world seek to lift lockdown measures without releasing a second wave of coronavirus. It is more nuanced than current health advice and more comprehensive than other privacy-preserving contact-tracing apps that have been proposed. Its efficacy does depend on a sense of social responsibility but it can also help breed that responsibility by heightening risk awareness.

Testing the network presents several challenges, and

mainly consists of running numerous scenarios to validate the model against what data there is:

- There are as yet very few general-population studies of COVID-19, meaning that almost all the knowledge we have so far is based on hospital patients.
- The studies that do exist are observational, not experimental.
- Little is known about how many people have COVID-19 without symptoms, as has been noted in Neil M, et al. [43].
- As soon as an observation is entered into the model, it no longer represents general population data.
- However, the first three of these problems are relevant for any predictive model. In the circumstances, a causal probabilistic model that uses expert knowledge is likely to be a better tool than other types of statistical analysis.

There are a few studies emerging that, while still not general-population surveys, do include people who are not ill or infected and provide valuable material for cross-checking probabilities. One such example is Poletti P, et al. [44], a study of 64,252 close contacts of 21,410 COVID-19 cases in Lombardy, Italy who were tested between February and April 2020. This study was aimed at discovering the probability of developing symptoms given infection, and was conditioned on age and sex. As testing becomes more prevalent and more such studies are possible, the Bayesian model can be further adjusted to reflect the latest state of knowledge.

Future Work

The revised complete BN model, which includes both the extended background risks and infection possibility described in this report together with the extended set of symptoms described in Butcher R, et al. [8], can be seen in the Appendix. More work needs to be done on immunity possibility, test types and results, and prevalence and infectiousness of asymptomatic and pre-symptomatic COVID-19 as more evidence become available. In general, the model needs to be constantly updated to keep up with the fast-developing state of scientific knowledge [45,46].

The power of the Bayesian model could be more explicitly leveraged by including interventions such as “go to school”, “wear a mask” or “take public transport”. If the network were to be used for policy-making rather than personal risk assessment, counterfactuals such as “close international borders”, “shut down pubs” or “postpone exams” could be added, taking advantage of another characteristic of Bayesian networks. Utility nodes could also be added - for example, to model the positive utility of recommencing education against the negative utility of increasing infection risks.

References

1. (2020) Comparisons of all- cause mortality between European countries and regions: January to June 2020. Office for National Statistics.
2. Smyth C, Wright O (2020) Failed coronavirus tracing app repackaged to tell users their risk ratings. The Times.
3. Funk S, Gilad E, Watkins C, Jansen VAA (2009) The spread of awareness and its impact on epidemic outbreaks. *Proc Natl Acad Sci USA* 106(16): 6782- 6877.
4. Dryhurst S, Schneider CR, Kerr J, Freeman AJ, Recchia G, et al. (2020) Risk perceptions of COVID-19 around the world. *Journal of Risk Research*.
5. Zoe (2020) COVID Symptom Study.
6. Allington D, Beaver K, Duffy B, Meyer C, Hall VM, et al. (2020) Getting used to life under lockdown? Coronavirus in the UK. King's College London.
7. Fenton N (2020) A note on 'Collider bias undermines our understanding of COVID-19 disease risk and severity' and how causal Bayesian networks both expose and resolve the problem. *arXiv*.
8. Butcher R, Fenton N (2020) Extending the range of symptoms in a Bayesian Network for the Predictive Diagnosis of COVID- 19. *MedRxiv*.
9. (2020) Coronavirus (COVID-19) Infection Survey UK Statistical bulletins. Office for National Statistics.
10. (2020) Centers for Disease Control and Prevention. Care for Children.
11. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, et al. (2020) Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 368(6491): 6936.
12. Kretzschmar ME, Rozhnova G, Bootsma MCJ, Boven M, Wiggert J, et al. (2020) Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet* 5(8): 452-459.
13. Huang Y, Sun M, Sui Y (2020) How Digital Contact Tracing Slowed Covid-19 in East Asia. *Harvard Business Review*.
14. Park S, Choi GJ, Ko H (2020) Information Technology-Based Tracing in Response to COVID-19 in South Korea-Privacy Controversies. *JAMA* 323(21): 2129-2130.
15. Asher S (2020) Coronavirus: Why Singapore turned to wearable contact-tracing tech.
16. Chan S (2020) COVID-19 Contact Tracing Apps Reach 9% Adoption In Most Populous Countries. *Sensor Tower*.
17. Hinch R, Probert W, Nurtay A, Kendall M, Wymant C, et al. (2020) Effective Configurations of a Digital Contact Tracing App: A report to NHSX.
18. Rasker R, Schunemann I, Barbar R, Vilcans K, Gray J, et al. (2020) Apps Gone Rogue: Maintaining Personal Privacy in an Epidemic. *ArXiv*.
19. Abate T (2020) Stanford researchers help develop privacy-focused coronavirus alert app. *Stanford*.
20. (2020) CoEpi: Community Epidemiology in Action. *CoEpi*.
21. Caramelo F, Ferreira N, Oliveiros B (2020) Estimation of risk factors for COVID-19 mortality-preliminary results. *MedRxiv*.
22. Evgeniou T, Fekom M, Ovchinnikov A, Porcher R, Pouchol C, et al. (2020) Personalised COVID-19 Isolation and Exit Policies Using Clinical Risk Predictions. *SSRN*.
23. Orlando V, Rea F, Savaré L, Guarino I, Mucherino S, et al. (2020) Development and validation of a clinical risk score to predict the risk of SARS-CoV-2 infection from administrative data: a population-based cohort study from Italy. *BioRxiv*.
24. Fenton N, McLachlan S, Lucas P, Dube K, Hitman G, et al. (2020) A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing. *MedRxiv*.
25. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, et al. (2020) Factors associated with COVID-19-related death using Open SAFELY. *Nature*.
26. (2020) Office for National Statistics.
27. (2020) Which occupations have the highest potential exposure to the coronavirus (COVID-19)? Office for National Statistics.
28. (2020) National Center for Health Statistics. *CDC*.
29. Takahashi T, Ellingson MK, Wong P, Israelow B, Lucas C, et al. (2020) Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*.
30. Pareek M, Bangash MN, Pareek N, Pan D, Sze S, et al. (2020) Ethnicity and COVID-19: an urgent public health research priority. *The Lancet* 395(10234): 1421-1422.
31. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, et al. (2020) Estimates of the severity of coronavirus

- disease 2019: a model-based analysis. *The Lancet* 20(6): 669-677.
32. Gebhard C, Zagrosek VR, Neuhauser HK, Morgan R, Klein SL (2020) Impact of sex and gender on COVID-19 outcomes in Europe. *Biol Sex Differ* 11: 29.
 33. Petrakis D, Margină D, Tsarouhas K, Tekos F, Stan M, et al. (2020) Obesity-a risk factor for increased COVID-19 prevalence, severity and lethality. *Molecular Medicine Reports* 22(1): 9-19.
 34. Clark A, Jit M, Gash CW, Guthrie B, Wang HX, et al. (2020) Global, regional and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *The Lancet* 8(8): 1003-1017.
 35. Blackshaw J (2020) Excess Weight and COVID-19 Insights from new evidence. London.
 36. Miyara M, Tubach F, Martinez V, Panzini CM, Pernet J, et al. (2020) Low rate of daily smokers in patients with symptomatic COVID-19. *MedRxiv*.
 37. Zyl Smit RN, Richards G, Leone FT (2020) Tobacco Smoking and COVID-19 infection. *The Lancet* 8(7): 664-665.
 38. Judah J (2020) After N.Y.C. Outbreak, Fearful British ultra-Orthodox Fight to Stave off Coronavirus. *Haaretz*.
 39. (2020) Why do some healthy people become seriously ill with COVID? Genomics England.
 40. Mackenzie D (2020) Race, COVID Mortality, and Simpson's Paradox. Los Angeles (California): UCLA.
 41. Emami A, Jayanmardi F, Pirbonyeh N, Akbari A (2020) Prevalence of Underlying Diseases in Hospitalized Patients with COVID-19: a Systematic Review and Meta-Analysis. *Arch Acad Emerg Med* 8(1): 35.
 42. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, et al. (2020) Features of 20 133 patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *The BMJ* 369: 1985.
 43. Neil M, Fenton N, Osman M, McLachlan S (2020) Bayesian network analysis of Covid-19 data reveals higher infection prevalence rates and lower fatality rates than widely reported. *Journal of Risk Research*.
 44. Poletti P, Tirani M, Cereda D, Trentini F, Guzzetta G, et al. (2020) Probability of symptoms and critical disease after SARS-CoV-2 infection. *ArXiv*.
 45. (2020) British Medical Association.
 46. Rhodes J, Subramanian S, Laird E, Griffin G, Kenny RA (2020) Prespective: Vitamin D deficiency and COVID-19 severity - plausibly linked by latitude, ethnicity, impacts on cytokines, ACE2 and thrombosis. *Journal of Internet Medicine*.

