

# A tool for COVID-19 containment using a Bayesian network model for personalised risk assessment

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

Supervisor: Professor Norman Fenton  
School of Electronic Engineering and Computer Science  
Queen Mary, University of London  
London, UK  
n.fenton@qmul.ac.uk

**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 a 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 behaviour 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.

**Keywords**—coronavirus, COVID-19, contact tracing, app, risk, privacy, second wave

## I. 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 (Office for National Statistics, 2020), even as business are encouraged to reopen, individuals urged to go back to work in offices, and schools reopen. Mitigating measures such as the compulsory wearing of face masks in confined public spaces have recently been imposed. But a digital contact-tracing system that was supposed to be the basis for a lifting of restrictions on movement and social activity has yet to be deployed nationally. The UK government is now trialling 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 (Smyth & Wright, 2020). 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 pan- and epidemics 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 (Funk, et al., 2009).

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, et al., 2020) 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 (Zoe, 2020). And, according to a poll of 2,254 UK residents carried out in May by King's College and Ipsos MORI (Allington, et al., 2020), the majority have 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, et al., 2020), 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.

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 Asian results in a raised probability of having severe COVID-19 in almost every age group. This is consistent with observed data (Office for National Statistics, 2020). (It is not 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 (Centers for Disease Control and Prevention, 2020). Adding the observation “multiple external interactions with other people” raises the probability of eventual COVID-19 by an order of magnitude.
- 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 overemphasising personal risks such as obesity or smoking.

## II. 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, et al., 2020) and (Kretzschmar, et al., 2020) suggest that the time from confirmation of a COVID-19 case to quarantining of contacts needs to be under 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 (Ferretti, et al., 2020) 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 (Huang, et al., 2020). 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 (Park, et al., 2020).

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 (Asher, 2020).

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 (Chan, 2020). 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 (Hinch, et al., 2020). 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 (Rasker, et al., 2020). 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 (Abate, 2020). 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 (CoEpi, 2020). 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 (Caramelo, et al., 2020). 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 (Evgeniou, et al., 2020). 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 (Orlando, et al., 2020). 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, et al., 2020).

### III. 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 Fig. 1, and a more detailed version can be found in the Appendix. The model, named “CombinedModelForContactTracing” 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 currently result in a 0.05% chance of eventually having COVID-19.

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 (The OpenSAFELY Collaborative, 2020) and (Office for National Statistics, 2020) 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 Fig. 2. All of these are factors that

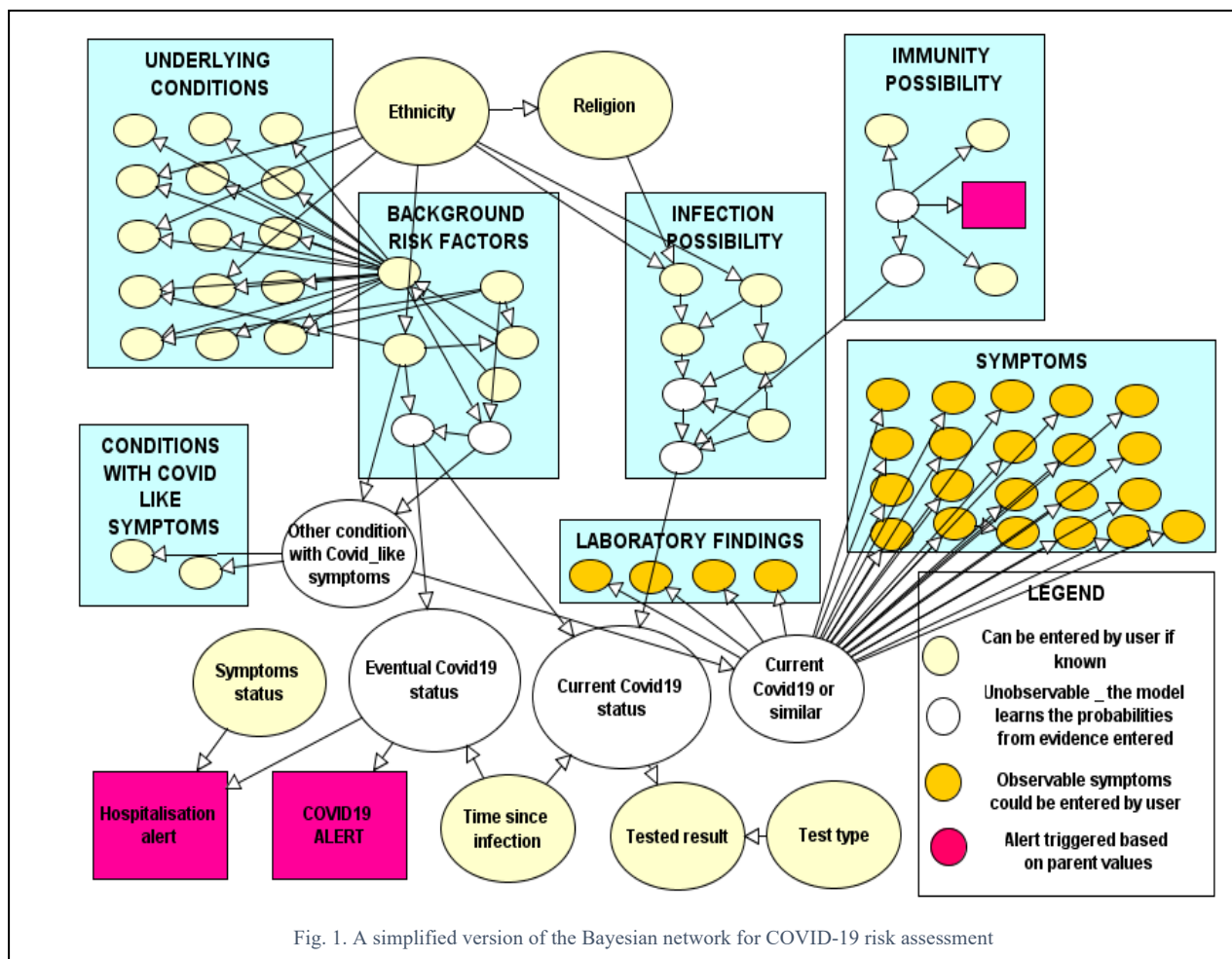
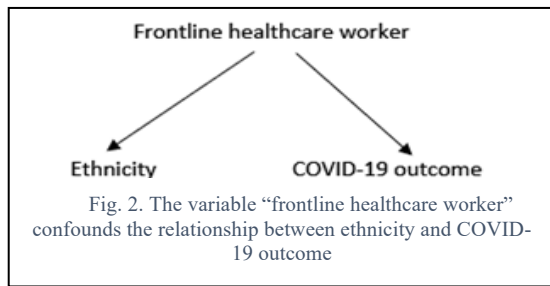


Fig. 1. A simplified version of the Bayesian network for COVID-19 risk assessment

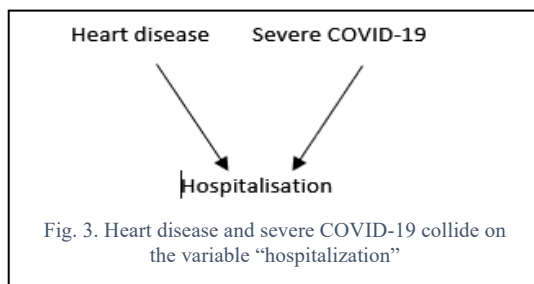


increase the risk of catching COVID-19 or may lead to a more severe form of the disease. They confound the relationship between ethnicity and COVID-19 outcomes.

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: hospital 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).

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 “hospitalised” we are confining our observations to a population already more likely to have heart disease or a whole range of other comorbidities, 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 Fig. 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 (National Center for Health Statistics, 2020).

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 (Takahashi, et al., 2020). 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.

#### IV. 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 Fig. 4. 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 Fig. 5.

The structure of the network, inherited from (Fenton, et al., 2020), 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 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 to 0.01% and of eventual severe COVID-19 from 0.01% to 0.02%. 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.

##### A. 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

Male	0.494
Female	0.506

Fig. 4. Node probability table for “sex”



Underlying medical condition	High						Medium						Low					
	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	0.725	0.834	1.0	1.0	1.0	1.0	1.0	1.0
True	0.166	0.166	0.153	0.326	0.275	0.166	0.166	0.166	0.153	0.326	0.275	0.166	0.0	0.0	0.0	0.0	0.0	0.0

Fig. 5. Node probability table for “diabetes”

dying of COVID-19 (Verity, et al., 2020) (Gebhard, et al., 2020) (Petrakis, et al., 2020) (Clark, et al., 2020) (Blackshaw, et al., 2020). 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 (Miyara, et al., 2020), this was soon dismissed by expert consensus, and the direct effects of smoking on COVID-19, if any, are still unknown (Zyl-Smit, et al., 2020) (Fenton, 2020). Smoking would therefore be included in the model as a parent node of “underlying medical condition” and also of chronic obstructive pulmonary disease (COPD).

### B. 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 socialise 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 (Judah, 2020).

## V. SELECTED PROBABILITY NODES

### A. 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, hospitalisation 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. (Genomics England, 2020). This is still in progress.

The media and political storm did cause a lot of data to be analysed by ethnicity and published, 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, et al., 2020) in Fig. 6.

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 be overall likely to make for higher risk in the absence of other knowledge. A fragment of

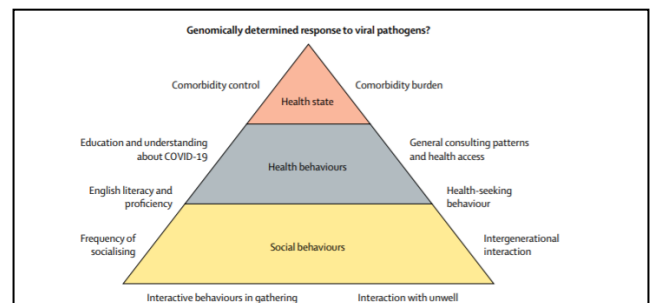


Fig. 6. The potential interaction of ethnicity-related factors on SARS-CoV-2 infection likelihood and COVID-19 outcomes (Pareek, et al., 2020)

the model showing the edges connecting ethnicity to other nodes is shown in Fig. 7. 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.

The connection between age and ethnicity is an important one. It explains the paradoxical situation that Black and non-Chinese Asian people have no increased risk of developing severe COVID-19 when considered as whole populations – in apparent contradiction of the news headlines – but they do have an 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 1.

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 (Mackenzie, 2020).

### B. 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 (The OpenSAFELY Collaborative, 2020) (Emami, et al., 2020) (Docherty, et al., 2020). 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

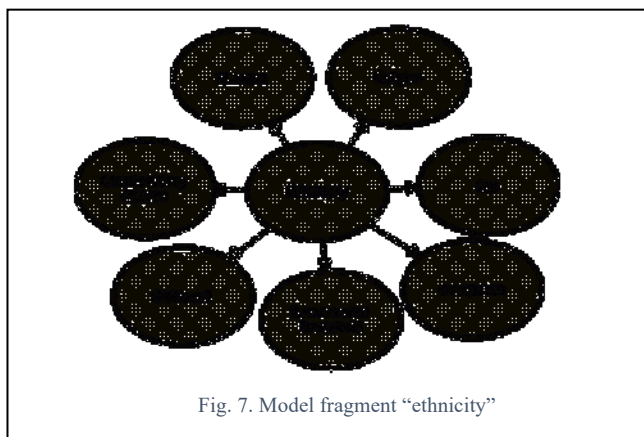


Fig. 7. Model fragment “ethnicity”

TABLE 1: PROBABILITY OF EVENTUAL SEVERE COVID-19 ILLNESS BY ETHNICITY AND AGE

AGE	White	Asian*	Black
Unobserved	0.01%	0.01%	0.01%
75+	0.02%	0.02%	0.02%
65-74	0.01%	0.02%	0.02%
55-64	0.01%	0.01%	0.02%
35-54	0.01%	0.01%	0.01%
25-34	0%	0.01%	0.01%
16-24	0%	0.01%	0.01%
<16	0%	0%	0.01%

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, hospitalisation, 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 2. (The data used differ slightly from those in the PHE report.)

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 3.

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.

TABLE 2. PROBABILITY OF OBESITY GIVEN SEVERE COVID-19

	Prior probability	Observe current COVID-19 status severe
Morbidly obese	2.846%	4.468%
Obese	23.353%	31.38%
Overweight	31.05%	34.012%
Normal/underweight	42.75%	30.14%

TABLE 3. PROBABILITY OF SEVERE COVID-19 GIVEN OBESITY

	Prior probability	Observe obese	Observe morbidly obese
COVID-19 status severe	0%	0%	0%
COVID-19 status mild	0.02%	0.02%	0.02%
COVID-19 status asymptomatic	0.03%	0.03%	0.03%
COVID-19 status none	99.95%	99.95%	99.95%

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.

### C. 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 (Office for National Statistics, 2020). 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.

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 becomes 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 for 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.

## VI. 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.

### A. 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 negligible and he is strongly predicted not to get COVID-19 at all, with a probability of 99.99%.

### B. 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. But the model also calculates her probability of eventual severe COVID-19 as negligible, with a 99.99% chance she will not become infected at all.

### C. 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 mild COVID-19 at 0.01%, higher than Fred’s or Katie’s but still below that of the general population. He has a 99.98% chance of not contracting COVID-19 at all.

### D. The African hospital orderly

Victor is a 29-year-old Oxford hospital orderly, Black, of Ghanaian origin. He is of normal weight, 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 flatmate who works from home. He has been in contact with patients with possible 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 1.07% chance of severe COVID-19 and a 6.75% chance of a mild form of the disease, well above the prior probabilities of 0.01% and 0.03% respectively, and only an 87.85% chance of not contracting the disease at all.

### E. 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 has only been to school a few times in the past months but her school is reopening this week. 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 probability of just 0.01% of having asymptomatic COVID-19 at the moment, the model predicts, with negligible probability of having it in a symptomatic form. Entering the observation “multiple external interactions with other people” to simulate the effect of her returning to school raises her probability of eventual mild symptomatic

COVID-19 to 0.02% and also doubles her probability of getting asymptomatic COVID-19. This is a risk that should be monitored closely and may need to be adjusted according to observational data once schools open.

## VII. 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, et al., 2020).
- 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, et al., 2020), 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.

## VIII. FUTURE WORK

After completing my sections of the model, background risks and infection possibility, I merged it with the symptoms section done by my colleague, Rachel Butcher. The complete model at this stage 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 becomes available. In general, the model needs to be constantly updated to keep up with the fast-developing state of scientific knowledge.

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”. It would be useful to include school or college attendance as a node in its own right.

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 to model, for example, the positive utility of recommencing education against the negative utility of increasing infection risks.

## REFERENCES

- Abate, T., 2020. *Stanford researchers help develop privacy-focused coronavirus alert app*. Stanford(California): s.n.
- Allington, D. et al., 2020. *Getting used to life under lockdown? Coronavirus in the UK*, London: s.n.
- Asher, S., 2020. *Coronavirus: Why Singapore turned to wearable contact-tracing tech*. [Online] Available at: <https://www.bbc.co.uk/news/technology-53146360> [Accessed 8 August 2020].
- Blackshaw, J. et al., 2020. *Excess Weight and COVID-19 Insights from new evidence*, London: s.n.
- Caramelo, F., Ferreira, N. & Oliveiros, B., 2020. *Estimation of risk factors for COVID-19 mortality - preliminary results*. s.l.:s.n.
- Centers for Disease Control and Prevention, 2020. *Care for Children*. [Online] Available at: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/pediatric-hcp.html#:~:text=A%20recent%20systematic%20review%20estimated,of%20pediatric%20infections%20are%20asymptomatic>. [Accessed 21 August 2020].
- Chan, S., 2020. *COVID-19 Contact Tracing Apps Reach 9% Adoption In Most Populous Countries*. [Online] Available at: <https://sensortower.com/blog/contact-tracing-app-adoption> [Accessed 21 August 2020].
- Clark, A. 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), pp. E1003-E1017.
- CoEpi, 2020. *CoEpi: Community Epidemiology in Action*. [Online] Available at: <https://www.coepi.org/> [Accessed 21 August 2020].
- Docherty, A. B. 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*, 22 May.
- Dryhurst, S. et al., 2020. Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, 5 May.
- 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. *Archives of Academic Emergency Medicine*, 24 March.

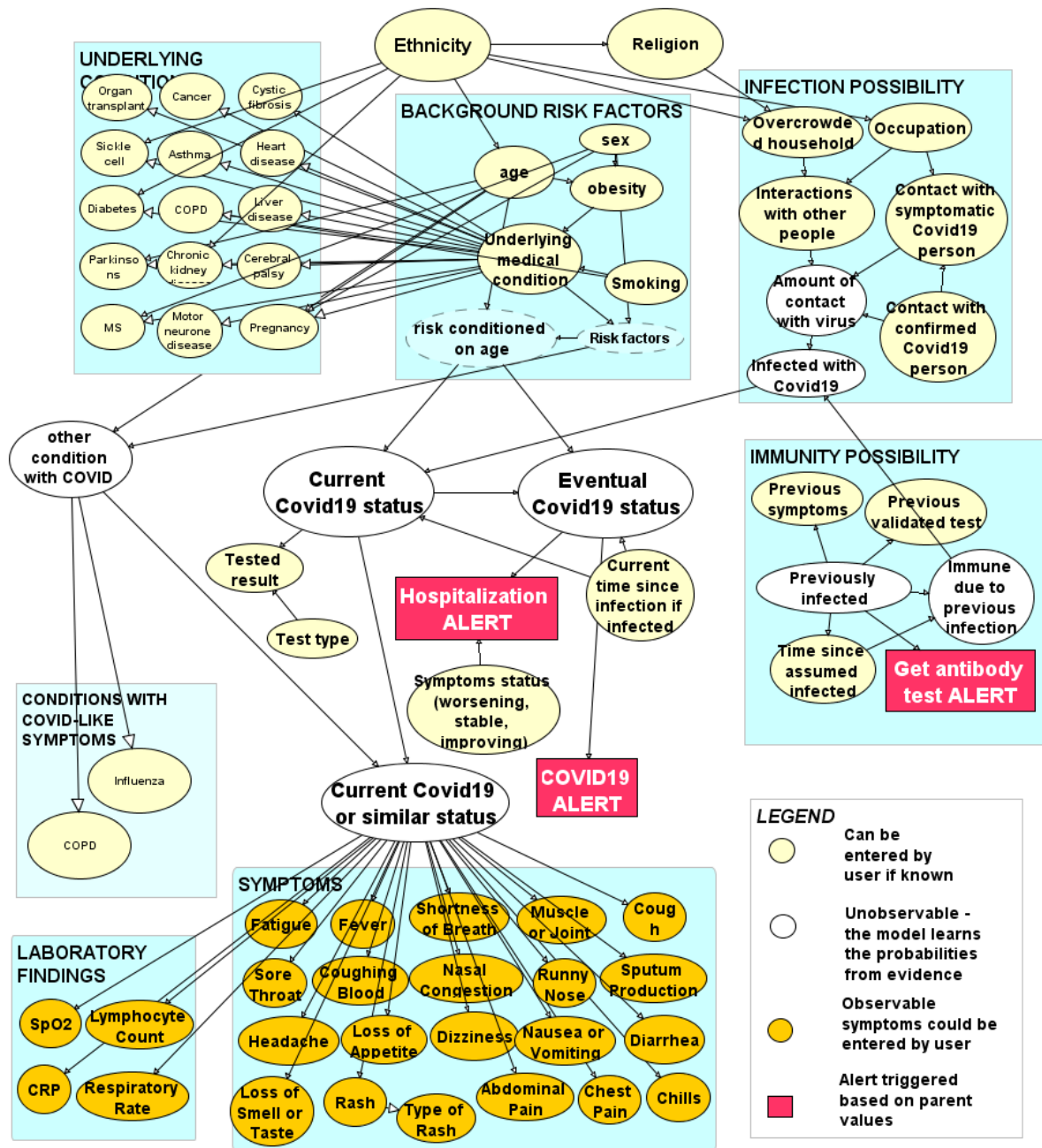


- Evgeniou, T. et al., 2020. *Personalised COVID-19 Isolation and Exit Policies Using Clinical Risk Predictions*. s.l.:s.n.
- Fenton, N., 2020. *A note on 'Collider bias undermines our understanding of COVID-19 disease risk and severity' and how causal Bayesian networks both expose and resolve the problem*, London: arXiv.
- Fenton, N. E. et al., 2020. *A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing*. London: medRxiv.
- Ferretti, L. et al., 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 8 May.368(6491).
- Funk, S., Gilad, E., Watkins, C. & Jansen, V. A. A., 2009. The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences of the United States of America*, 21 April, 106(16), pp. 6782-6877.
- Gebhard, C. et al., 2020. Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of Sex Differences*, 25 May. Volume 11.
- Genomics England, 2020. *Genomics England*. [Online] Available at: <https://www.genomicsengland.co.uk/genomics-england-genomicc-nhs-covid-19/> [Accessed 9 August 2020].
- Hinch, R. et al., 2020. *Effective Configurations of a Digital Contact Tracing App: A report to NHSX*, s.l.: s.n.
- Huang, Y., Sun, M. & Sui, Y., 2020. How Digital Contact Tracing Slowed Covid-19 in East Asia. *Harvard Business Review*, 15 April.
- Judah, J., 2020. *After N.Y.C. Outbreak, Fearful British ultra-Orthodox Fight to Stave Off Coronavirus*, London: Haaretz.
- Kretzschmar, M. E. et al., 2020. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet*, 16 July, 5(8), pp. E452-E459.
- Mackenzie, D., 2020. *Race, COVID Mortality, and Simpson's Paradox*. Los Angeles(California): UCLA.
- Miyara, M. et al., 2020. *Low rate of daily smokers in patients with symptomatic COVID-19*, Paris: medRxiv.
- National Center for Health Statistics, 2020. *CDC*. [Online] Available at: [https://www.cdc.gov/nchs/nvss/vsrr/covid\\_weekly/index.htm?fbclid=IwAR3-wrg3tTKK5-9tOHPGAHWFO3DfslkJ0KsDEPQpWmPbKtp6EsoVV2Qs1Q#ExcessDeaths](https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm?fbclid=IwAR3-wrg3tTKK5-9tOHPGAHWFO3DfslkJ0KsDEPQpWmPbKtp6EsoVV2Qs1Q#ExcessDeaths) [Accessed 1 September 2020].
- Neil, Martin., Fenton, Norman., Osman, Magda. & McLachlan, Scott., 2020. Bayesian network analysis of Covid-19 data reveals higher infection prevalence rates and lower fatality rates than widely reported. *Journal of Risk Research*, 29 June.
- Office for National Statistics, 2020. *Comparisons of all-cause mortality between European countries and regions: January to June 2020*, London: UK government.
- Office for National Statistics, 2020. *ons.gov.uk*. [Online] Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronaviruscovid19relateddeathsbyethnicgroupenglandandwales/2march2020to15may2020> [Accessed 20 August 2020].
- Office for National Statistics, 2020. *Which occupations have the highest potential exposure to the coronavirus (COVID-19)?*. [Online] Available at: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/whichoccupationshavethehighestpotentialexposuretothecoronaviruscovid19/2020-05-11> [Accessed 9 August 2020].
- Orlando, V. 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*. s.l.:s.n.
- Pareek, M. et al., 2020. Ethnicity and COVID-19: an urgent public health research priority. *The Lancet*, 395(10234), pp. 1421-1422.
- Park, S., Choi, G. J. & Ko, H., 2020. Information Technology-Based Tracing in Response to COVID-19 in South Korea - Privacy Controversies. *JAMA Network*, 23 April.
- Petrakis, D. et al., 2020. Obesity - a risk factor for increased COVID-19 prevalence, severity and lethality. *Molecular Medicine Reports*, 22(1), pp. 9-19.
- Poletti, P. et al., 2020. *Probability of symptoms and critical disease after SARS-CoV-2 infection*, Ithaca: s.n.
- Rasker, R. et al., 2020. *Apps Gone Rogue: Maintaining Personal Privacy in an Epidemic*. Boston(Massachusetts): arXiv.org.
- Smyth, C. & Wright, O., 2020. [Online] Available at: <https://www.thetimes.co.uk/article/coronavirus-failed-tracing-app-repackaged-to-tell-users-their-risk-ratings-l88n3zhnx> [Accessed 6 August 2020].
- Takahashi, T. et al., 2020. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*, 26 August.
- The OpenSAFELY Collaborative, 2020. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 1 July.
- Verity, R. et al., 2020. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet*, 1 June, 20(6), pp. 669-677.
- Zoe, 2020. *COVID Symptom Study*. [Online] Available at: <https://covid.joinzoe.com/> [Accessed 30 August 2020].
- Zyl-Smit, R. N. v., Richards, G. & Leone, F. T., 2020. Tobacco Smoking and COVID-19 infection. *The Lancet*, 25 May, 8(7), pp. 664-665.

## APPENDIX

This appendix shows the full model, followed by the nodes on which I worked and details how I arrived at their probability tables.

### A. FULL MODEL



The model, entitled “CombinedModelForContactTracing.cmpx” can be downloaded at Norman Fenton’s home page <http://www.eecs.qmul.ac.uk/~norman/Models/> and can be run using trial software from agenarisk.com.

### B. NODES

The nodes are presented in the following sections:

1. Ethnicity
2. Religion

3. Other major background risk factors
4. Underlying medical conditions
5. Risk factors (combined)
6. Risk conditioned on age
7. Infection risks
8. Outcomes

Other nodes were either inherited from an earlier version of the model or were worked on by my colleague Rachel Butcher.

## 1. ETHNICITY

White	0.84815186
Mixed/Multiple	0.017982017
Chinese	0.006993007
Other Asian	0.07292707
Black	0.034965035
Other	0.018981019

These data were taken from the latest population estimates (2019) by the Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/researchreportonpopulationestimatesbyethnicgroupandreligion/2019-12-04>.

They are only for England and Wales but I decided that extrapolating the proportions to the rest of the UK was preferable to taking the 2011 Census statistics for the whole of the UK.

I supplemented this with government statistics on the Chinese ethnic group. <https://www.ethnicity-facts-figures.service.gov.uk/summaries/chinese-ethnic-group>.

It seemed important to separate Chinese from other Asian groups because of the very different health and cultural profiles of Chinese and South Asian.

It would also have been useful to separate Black African and Black Caribbean, for the same reasons.

## 2. RELIGION

Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
Christian	0.6392776	0.4625637	0.19573641	0.09970134	0.6908563	0.19763666
Buddhist	0.0017348279	0.00804884	0.12551223	0.025770145	0.001506255	0.0064715734
Hindu	2.4945344E-4	0.007972068	0.0073408782	0.20372605	0.0029352938	0.01449895
Jewish	0.0050467337	0.0034702711	8.2921906E-4	6.392018E-4	8.638579E-4	0.020255601
Muslim	0.004368858	0.083781436	0.02041761	0.47705418	0.14586115	0.51497436
Sikh	1.5474162E-4	0.0041832733	0.0025919455	0.09619018	7.673375E-4	0.07209915
Other	0.003790608	0.0060160076	0.0039553237	0.009977516	0.0038066588	0.0064733475
None	0.2730012	0.3232996	0.55641866	0.034573957	0.0737132	0.0919893
Not stated	0.07237602	0.10066482	0.08719772	0.05236743	0.07968996	0.075601034

These data came from the ONS via a 2015 Freedom of Information request.

<https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/ethnicityandreligionbyage>

The absolute numbers in the xls file are normalised by the AgenaRisk software to add up to 1 in each column.

## 3. BACKGROUND RISK FACTORS

### 3.a. AGE

Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
75 and above	0.08727237	0.012266416	0.020000458	0.021542113	0.025700176	0.017351551
65-74	0.095409095	0.016625285	0.029999414	0.035962854	0.036260583	0.028616488
55-64	0.12686937	0.028467821	0.07199962	0.06665908	0.04970856	0.0607597
35-54	0.2798486	0.16767396	0.24899973	0.2605109	0.3147982	0.2778235
25-34	0.124623135	0.14566073	0.21999995	0.21280235	0.16364022	0.22974086
16-24	0.11281764	0.1795067	0.29099992	0.14314783	0.13706492	0.14736845
0-15	0.1731598	0.4497991	0.11800092	0.2593749	0.27282736	0.23833945

I used data from the same FOI request as for the Religion node above.

The numbers needed some rearranging into the age groupings I had chosen. Here and in other similar situations I used this population-distribution calculator from the ONS to weight the results:

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/august2019>

### 3.b. SEX

Male	0.494
Female	0.506

Data

source:

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates>

### 3.c. OBESITY

age	75 and above		65-74		55-64		35-54	
sex	Male	Female	Male	Female	Male	Female	Male	Female
morbidly obese	0.009	0.02	0.022	0.02	0.031	0.05	0.036	0.056
obese	0.286	0.3	0.298	0.31	0.325	0.27	0.287	0.271
overweight	0.458	0.38	0.46	0.4	0.428	0.35	0.428	0.3
normal/unde...	0.247	0.3	0.22	0.27	0.216	0.33	0.249	0.373

age	25-34		16-24		0-15	
sex	Male	Female	Male	Female	Male	Female
morbidly obese	0.021	0.05	0.003	0.05	0.0	0.0
obese	0.174	0.21	0.103	0.16	0.17	0.16
overweight	0.382	0.29	0.222	0.22	0.12	0.13
normal/unde...	0.423	0.45	0.672	0.57	0.71	0.71

These data were taken from the NHS Health Survey for England 2017. (I could not find such detailed data in more recent surveys.)

<https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2017>

There were separate government figures available on obesity by ethnicity:

<https://www.ethnicity-facts-figures.service.gov.uk/health/diet-and-exercise/overweight-adults/latest>

But I was unable to find data on obesity by all three factors. I decided in favour of age and sex because a quick Excel analysis suggested they were bigger influencers of obesity than ethnicity. I was also able to capture some of the relationship between ethnicity and obesity by conditioning diabetes on ethnicity (see Diabetes section below).

### 3.d. SMOKING

current	0.134
previous	0.352
never	0.514

Data

are

from

here:

<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2019#data-on-smokers-who-have-quit-and-smokers-who-intend-to-quit-great-britain-1974-to-2019>

The numbers are for adults aged 16 and above. I decided to assume there were no smokers aged below 16, given that numbers would be small and risks at that age, in terms of indicating an underlying health condition, would be minimal.

### 3.e. UNDERLYING MEDICAL CONDITION

Smoking	current				previous				never			
obesity	morbid...	obese	overw...	normal...	morbid...	obese	overw...	normal...	morbid...	obese	overw...	normal...
High	0.8	0.8	0.1	0.1	0.8	0.1	0.1	0.1	0.8	0.1	0.1	0.1
Medium	0.1	0.1	0.8	0.1	0.1	0.8	0.8	0.1	0.1	0.8	0.1	0.1
Low	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.8	0.1	0.1	0.8	0.8

These are rough approximations conditioned on smoking and obesity status. The probabilities are altered by backward inference when an observation is made for any specific medical condition.

### 4. UNDERLYING CONDITIONS:

I took the risk classification from here: <https://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/whos-at-higher-risk-from-coronavirus/>

For each underlying condition, I first took the classification of high- or medium-risk, summed the total high-risk and medium-risk patients, and then calculated how probable this particular condition was given an overall rating of high or medium for “underlying medical condition”.

HIGH RISK	organ	cancer	cystic	sickle	TOTAL HIGH								
number of cases	53000	2900000	10500	15000	2978500								
% of population	0%	4%	0%	-									
% of high risk	1.80%	97.30%	0.40%	0.50%									
% of all underlying													
MODERATE RISK	asthma	COPD	heart	diabetes	kidney	liver	Parkinson's	motor	MS	cerebral	pregnant	TOTAL MODERATE	
number of cases	5400000	1500000	2300000	4700000	3000000	9300	145000	5000	130000	30000	549000	17768300	
% of population	8%	2%	4%	7%	5%	0%	0%	0%	0%	0%	1%		
% of moderate risk	30.39%	8.44%	12.94%	26.45%	16.88%	0.05%	0.82%	0.03%	0.73%	0.17%	3.09%		
% of all underlying	26%	7.23%	11.10%	22.70%	14.50%	0.04%	0.70%	0.02%	0.63%	0.14%	2.65%		

#### 4.a. ORGAN TRANSPLANT

Underlying...	High	Medium	Low
False	0.982	1.0	1.0
True	0.018	0.0	0.0

<https://www.organdonation.nhs.uk/get-involved/news/more-than-50-000-now-alive-thanks-to-organ-donations/>

#### 4.b. CANCER

Underlying...	High	Medium	Low
False	0.027	1.0	1.0
True	0.973	0.0	0.0

[https://www.macmillan.org.uk/\\_images/cancer-statistics-factsheet\\_tcm9-260514.pdf](https://www.macmillan.org.uk/_images/cancer-statistics-factsheet_tcm9-260514.pdf)

#### 4.c. CYSTIC FIBROSIS



Underlying...	High	Medium	Low
False	0.996	1.0	1.0
True	0.004	0.0	0.0

<https://www.cysticfibrosis.org.uk/what-is-cystic-fibrosis/faqs#:~:text=Around%2010%2C500%20people%20in%20the,100%2C000%20people%20in%20the%20world.>

#### 4.d. SICKLE CELL DISEASE

Underlying ...	High					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	0.99822	0.99822	0.99822	0.99822	0.91	0.99822
True	0.00178	0.00178	0.00178	0.00178	0.09	0.00178

Underlying ...	Medium					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0

Underlying ...	Low					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0

<https://www.sicklecellsociety.org/about-sickle-cell/#:~:text=4%20Approximately%2015%2C000%20people%20in,in%20the%20UK%20every%20year.&text=7%20Childr en%20with%20SCD%20are,ages%20of%202%20and%2016>

#### 4.e. ASTHMA

Underlying...	High	Medium	Low
False	0.74	0.74	1.0
True	0.26	0.26	0.0

<https://www.asthma.org.uk/about/media/facts-and-statistics/>

#### 4.f. HEART DISEASE

Underlying...	High	Medium	Low
False	0.889	0.889	1.0
True	0.111	0.111	0.0

<https://www.bhf.org.uk/what-we-do/our-research/heart-statistics>

#### 4.g. DIABETES

Underlying ...	High					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	0.834	0.834	0.847	0.674	0.725	0.834
True	0.166	0.166	0.153	0.326	0.275	0.166

Underlying ...	Medium					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	0.834	0.834	0.847	0.674	0.725	0.834
True	0.166	0.166	0.153	0.326	0.275	0.166

Underlying ...	Low					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0

[https://www.diabetes.org.uk/resources-s3/2017-11/diabetes\\_in\\_the\\_uk\\_2010.pdf](https://www.diabetes.org.uk/resources-s3/2017-11/diabetes_in_the_uk_2010.pdf)

#### 4.h. COPD

Underlying ...	High			Medium			Low		
	current	previous	never	current	previous	never	current	previous	never
False	0.935	0.935	0.993	0.935	0.935	0.993	1.0	1.0	1.0
True	0.065	0.065	0.0069999998	0.065	0.065	0.0069999998	0.0	0.0	0.0

<https://www.hse.gov.uk/statistics/causdis/copd.pdf>

<https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/causes/>

#### 4.i. LIVER DISEASE

Underlying...	High	Medium	Low
False	0.9996	0.9996	1.0
True	4.0E-4	4.0E-4	0.0

<https://www.statista.com/statistics/1036415/prevalence-of-cirrhosis-in-the-uk-by-gender/>

#### 4.j. PARKINSON'S DISEASE

Underlying ...	High						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	0.974	0.9931	0.9994	0.99998	1.0	1.0	1.0
True	0.025999999	0.0069	5.9999997E-4	2.0000001E-5	0.0	0.0	0.0

Medium						
75 and above	65-74	55-64	35-54	25-34	16-24	0-15
0.974	0.9931	0.9994	0.99998	1.0	1.0	1.0
0.025999999	0.0069	5.9999997E-4	2.0000001E-5	0.0	0.0	0.0

Low						
75 and above	65-74	55-64	35-54	25-34	16-24	0-15
1.0	1.0	1.0	1.0	1.0	1.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

<https://www.parkinsons.org.uk/professionals/news/parkinsons-diagnoses-rise-uk#:~:text=The%20number%20of%20people%20diagnosed,350%20adults%20in%20the%20UK>

#### 4.k. CHRONIC KIDNEY DISEASE

Underlying ...	High					
Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
False	0.862	0.938	0.876	0.8	0.736	0.938
True	0.138	0.062	0.124	0.2	0.264	0.062

Medium					
White	Mixed/Multiple	Chinese	Other Asian	Black	Other
0.862	0.938	0.876	0.8	0.736	0.938
0.138	0.062	0.124	0.2	0.264	0.062

Low					
White	Mixed/Multiple	Chinese	Other Asian	Black	Other
1.0	1.0	1.0	1.0	1.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0

<https://www.renalreg.org/wp-content/uploads/2014/09/06-Chap-06.pdf>

#### 4.1. CEREBRAL PALSY

Underlying...	High	Medium	Low
False	0.9986	0.9986	1.0
True	0.0014	0.0014	0.0

<https://thepacecentre.org/information-centre/stats-facts/>

#### 4.m. MULTIPLE SCLEROSIS

Underlying ...	High		Medium		Low	
sex	Male	Female	Male	Female	Male	Female
False	0.9989	0.9973	0.9989	0.9973	1.0	1.0
True	0.0011	0.0027	0.0011	0.0027	0.0	0.0

<https://www.gov.uk/government/publications/multiple-sclerosis-prevalence-incidence-and-smoking-status/multiple-sclerosis-prevalence-incidence-and-smoking-status-data-briefing>

#### 4.n. MOTOR NEURONE DISEASE

Underlying...	High	Medium	Low
False	0.9998	0.9998	1.0
True	2.0E-4	2.0E-4	0.0

<https://www.nhsinform.scot/illnesses-and-conditions/brain-nerves-and-spinal-cord/motor-neurone-disease-mnd#:~:text=Motor%20neurone%20disease%20is%20a%20rare%20condition%20that%20affects%20around,although%20th is%20is%20extremely%20rare.>

#### 4.o. PREGNANCY

Underlying ...	High						
sex	Male						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Underlying ...	High						
sex	Female						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	0.994006	0.818	0.911	1.0
True	0.0	0.0	0.0	0.005994006	0.182	0.089	0.0

Underlying ...	Medium						
sex	Male						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0	0.0


Underlying ...	Medium						
sex	Female						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	0.994	0.818	0.911	1.0
True	0.0	0.0	0.0	0.006	0.182	0.089	0.0

Underlying ...	Low						
sex	Male						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0	0.0


Underlying ...	Low						
sex	Female						
age	75 and above	65-74	55-64	35-54	25-34	16-24	0-15
False	1.0	1.0	1.0	1.0	1.0	1.0	1.0
True	0.0	0.0	0.0	0.0	0.0	0.0	0.0

<https://www.statista.com/statistics/297718/conception-rate-per-thousand-women-in-england-and-wales-by-age/>

#### 5. RISK FACTORS

Expression Type	 <b>TNormal</b> Normal distribution truncated at finite end values
Mean	wmean(10.0,underlying,1.0,sex)
Variance	1.0E-4

## 6. RISK CONDITIONED ON AGE

Expression Type	 <b>TNormal</b> Normal distribution truncated at finite end values
Mean	wmax(10.0,age,1.0,risks)
Variance	0.01

## 7. INFECTION RISKS

### 7.a. OCCUPATION

Ethnicity	White	Mixed/Multiple	Chinese	Other Asian	Black	Other
Frontline he...	0.026	0.017	0.006	0.039	0.051	0.059
Other essential	0.304	0.313	0.324	0.291	0.279	0.271
Non-essenti...	0.67	0.67	0.67	0.67	0.67	0.67

<https://www.ethnicity-facts-figures.service.gov.uk/workforce-and-business/workforce-diversity/nhs-workforce/latest>

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/908434/Disparities\\_in\\_the\\_risk\\_and\\_outcomes\\_of\\_COVID\\_August\\_2020\\_update.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908434/Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf)



## 7.b. OVERCROWDED HOUSEHOLD

Ethnicity	White									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.02	0.02	0.02	0.09	0.02	0.02	0.02	0.02	0.02	0.02
No	0.98	0.98	0.98	0.91	0.98	0.98	0.98	0.98	0.98	0.98

Ethnicity	Mixed/Multiple									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.058000002	0.058000002	0.058000002	0.09	0.058000002	0.058000002	0.058000002	0.058000002	0.058000002	0.058000002
No	0.942	0.942	0.942	0.91	0.942	0.942	0.942	0.942	0.942	0.942

Ethnicity	Chinese									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.07	0.07	0.07	0.09	0.07	0.07	0.07	0.07	0.07	0.07
No	0.93	0.93	0.93	0.91	0.93	0.93	0.93	0.93	0.93	0.93

Ethnicity	Other Asian									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.124	0.124	0.124	0.09	0.124	0.124	0.124	0.124	0.124	0.124
No	0.876	0.876	0.876	0.91	0.876	0.876	0.876	0.876	0.876	0.876

Ethnicity	Black									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.121999994	0.121999994	0.121999994	0.09	0.121999994	0.121999994	0.121999994	0.121999994	0.121999994	0.121999994
No	0.878	0.878	0.878	0.91	0.878	0.878	0.878	0.878	0.878	0.878

Ethnicity	Other									
Religion	Christian	Buddhist	Hindu	Jewish	Muslim	Sikh	Other	None	Not stated	
Yes	0.084	0.084	0.084	0.09	0.084	0.084	0.084	0.084	0.084	0.084
No	0.916	0.916	0.916	0.91	0.916	0.916	0.916	0.916	0.916	0.916

<https://www.ethnicity-facts-figures.service.gov.uk/housing/housing-conditions/overcrowded-households/latest>

[https://www.jpr.org.uk/documents/JPR\\_Census\\_Jewish\\_families\\_and\\_Jewish\\_households\\_report\\_March\\_2015.pdf](https://www.jpr.org.uk/documents/JPR_Census_Jewish_families_and_Jewish_households_report_March_2015.pdf)

## 7.c. CONTACT WITH SYMPTOMATIC COVID-19 PERSON

Contact wit...	Yes			No		
Occupation	Frontline healthc...	Other essential	Non-essential/no...	Frontline healthc...	Other essential	Non-essential/no...
Yes	0.9	0.8	0.8	0.1	5.0226017E-4	1.00090074E-4
No	0.1	0.2	0.2	0.9	0.9994977	0.9998999

## 7.d. CONTACT WITH CONFIRMED COVID-19 PERSON


Yes	8.0000005E-5
No	0.99992

<https://www.gov.uk/government/publications/nhs-test-and-trace-england-and-coronavirus-testing-uk-statistics-6-august-to-12-august-2020/weekly-statistics-for-nhs-test-and-trace-england-and-coronavirus-testing-uk-6-august-to-12-august>

## 7.e. INTERACTIONS WITH OTHER PEOPLE

Occupation	Frontline healthcare		Other essential		Non-essential/not working	
Overcrowd...	Yes	No	Yes	No	Yes	No
Multiple exte...	1.0	0.9	1.0	0.3	0.9	0.0057803467
Some external	0.0	0.1	0.0	0.7	0.1	0.0115606915
Minimal	0.0	0.0	0.0	0.0	0.0	0.982659

## 7.f. AMOUNT OF CONTACT WITH VIRUS

Expression Type	 <b>TNormal</b> Normal distribution truncated at finite end values
Mean	wmean(5.0,recent_contact_symptoms,1.0,multiple_interactions,5.0,recent_contact_confirmed)
Variance	0.001

## 8. OUTCOMES

### 8.a. INFECTED WITH COVID-19

Immune du...	False				
Amount of ...	Very High	High	Medium	Low	None
Yes	0.9	0.5	2.0E-4	1.0000001E-5	0.0
No	0.1	0.5	0.9998	0.99999	1.0

Immune du...	True				
Amount of ...	Very High	High	Medium	Low	None
Yes	0.0	0.0	0.0	0.0	0.0
No	1.0	1.0	1.0	1.0	1.0

Government community infection survey pilot:

<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronavirus/covid19infectionsurveysurvey/pilot/englandandwales21august2020>

### 8.b. CURRENT TIME SINCE INFECTION IF INFECTED

> 5 days	0.5
<= 5 days	0.5

This on how long COVID-19 remains infectious: <https://covid.joinzoe.com/post/covid-long-term>

And this on incubation period: <https://www.acpjournals.org/doi/10.7326/M20-0504>

### 8.c. CURRENT COVID-19 STATUS

Infected wi...	Yes					
risk conditio...	High		Medium		Low	
Current tim...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 sev...	0.4	2.0090406E-4	0.1	3.0089366E-5	0.05	0.0
Covid19 mild	0.5	3.013561E-4	0.7	3.0089365E-4	0.8	0.001
Covid19 asy...	0.1	0.9994977	0.2	0.999669	0.15	0.999
None	0.0	0.0	0.0	0.0	0.0	0.0

Infected wi...	No					
risk conditio...	High		Medium		Low	
Current tim...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 sev...	0.0	0.0	0.0	0.0	0.0	0.0
Covid19 mild	0.0	0.0	0.0	0.0	0.0	0.0
Covid19 asy...	0.0	0.0	0.0	0.0	0.0	0.0
None	1.0	1.0	1.0	1.0	1.0	1.0

<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronavirus/covid19infectionsurveysurvey/pilot/englandandwales21august2020>

#### 8.d. EVENTUAL COVID-19 STATUS

Current Covid19 status	Covid19 severe					
risk conditioned on age	High		Medium		Low	
Current time since inf...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 severe	1.0	1.0	1.0	1.0	1.0	1.0
Covid19 mild	0.0	0.0	0.0	0.0	0.0	0.0
Covid19 asymptomatic	0.0	0.0	0.0	0.0	0.0	0.0
None	0.0	0.0	0.0	0.0	0.0	0.0

Current Covid19 status	Covid19 mild					
risk conditioned on age	High		Medium		Low	
Current time since inf...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 severe	0.2	0.9	0.1	0.5	0.01	0.1
Covid19 mild	0.8	0.1	0.9	0.5	0.99	0.9
Covid19 asymptomatic	0.0	0.0	0.0	0.0	0.0	0.0
None	0.0	0.0	0.0	0.0	0.0	0.0

Current Covid19 status	Covid19 asymptomatic					
risk conditioned on age	High		Medium		Low	
Current time since inf...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 severe	0.1	0.7	0.01	0.1	0.0	0.05
Covid19 mild	0.5	0.25	0.29	0.7	0.2	0.1
Covid19 asymptomatic	0.4	0.05	0.7	0.2	0.8	0.85
None	0.0	0.0	0.0	0.0	0.0	0.0

Current Covid19 status	None					
risk conditioned on age	High		Medium		Low	
Current time since inf...	> 5 days	<= 5 days	> 5 days	<= 5 days	> 5 days	<= 5 days
Covid19 severe	0.0	0.0	0.0	0.0	0.0	0.0
Covid19 mild	0.0	0.0	0.0	0.0	0.0	0.0
Covid19 asymptomatic	0.0	0.0	0.0	0.0	0.0	0.0
None	1.0	1.0	1.0	1.0	1.0	1.0

# MSc Project - Reflective Essay

<b>Project Title:</b>	A tool for COVID-19 containment using a Bayesian network model for personalised risk assessment
<b>Supervisor Name:</b>	Norman Fenton
<b>Programme of Study:</b>	MSc Computing and Information Systems

## INTRODUCTION

This project grew out of an emerging body of work on the risks of COVID-19 by Norman Fenton, Scott McLachlan and a group of their colleagues (Fenton, et al., 2020). The work modelled the probability of novel coronavirus infection and illness using a Bayesian causal network that captured a multitude of interrelated factors to allow calculations of individual risk. At the time I began work on the project, Britain was three months into a lockdown, and a blizzard of statistics being released about deaths, illness, risks, permitted behaviours and government policy was causing alarm and confusion among the public. I was attracted to this work because of the unique and elegant way Bayesian networks can model complex relationships, providing a powerful aid to reasoning in an environment of uncertainty. They are a tool for both predicting the effects of interventions and for diagnosing the causes of observed data. Picking up where Professor Fenton and his colleagues had left off, I augmented the model with details of background risks, infection risks and how these influenced eventual outcomes, and finally recalibrated the probabilities of COVID infection and degree of illness according to the latest research available.

## STRENGTHS AND WEAKNESSES

The principal strength of a Bayesian approach in general and this project in particular is their ability to model complex relationships in a causal way that goes beyond simple correlation or association (Jensen, 2009). The ability to work with a high level of uncertainty is vital for a situation such as the current one in which a new, deadly and highly contagious new virus necessitates urgent action in an environment of relative ignorance. Bayesian networks – essentially joint probability tables - work well with even minimal data, are capable of modelling interventions and counterfactuals as well as observations, and are highly adaptable to new data and expert knowledge (Pearl & MacKenzie, 2018).

Building a causal network of risk factors provided a method that was both visually intuitive and enforced a discipline of thought about which factors influenced which others, and to what degree. A snippet of the network is depicted in Fig. 1 below. It helped answer questions such as:

- What relationship does ethnicity bear to eventual severity of COVID-19, and why?
- Is age or underlying medical condition a more important determinant of how seriously ill an infected person becomes, and how are those related to one another?
- How risky is it to have multiple interactions with other people, even if they show no signs of having coronavirus?
- How likely is an asymptomatic person infected with COVID-19 likely to remain asymptomatic?
- Does smoking have any direct effect on the severity of COVID-19?

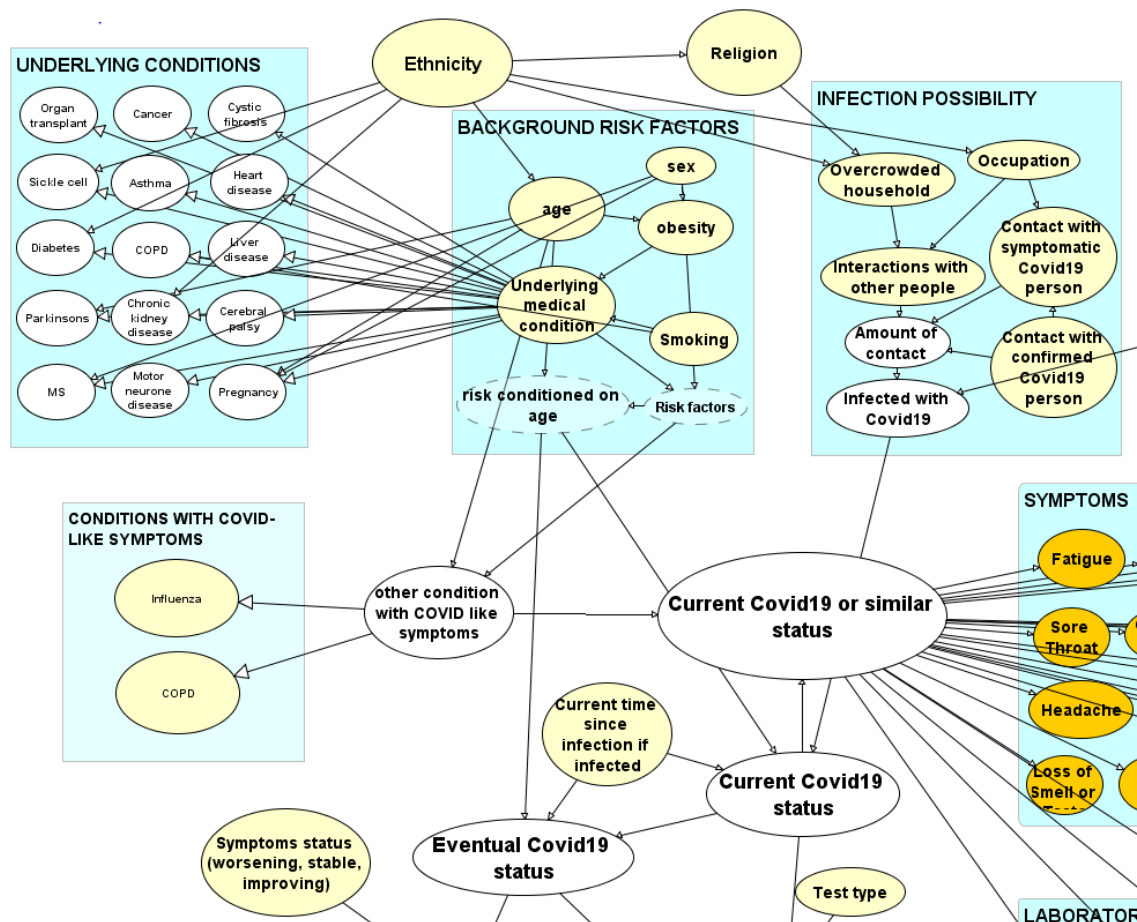


Fig. 1: A snippet of the Bayesian network for modelling COVID-19 risk assessment showing the relationships between background risk factors, infection possibilities and outcomes

When I began my research in June 2020 there were almost no data available about the spread of COVID-19 in the community. Very little testing had been done outside hospitals, and practically none on people without COVID-19 symptoms. This made the assignment of prior probabilities of vulnerability in the general population challenging. It seemed likely from early studies in China that older age was likely the most important risk factor, and that being male was also a significant risk factor (Verity, et al., 2020). Notably, however, there was conjecture early in the pandemic that the higher prevalence and severity in men might be due to the fact that many more Chinese men than women smoke (Cai, 2020). This was one of many theories that were floated and subsequently side-lined as experts struggled to understand how the novel coronavirus worked. Later, another theory briefly surfaced that smoking had a protective effect (Miyara, et al., 2020). A Bayesian network proved a nimble tool to represent these changing theories, as lines representing causal relationships could be redrawn or deleted as new evidence emerged, while basic data about the prevalence of, e.g., smoking in the general population could be preserved.

More importantly, thinking about associations in a causal way was a sharp tool for analysing certain relationships that only existed due to a common cause or effect – confounders or colliders – or through mediating factors.



To take an example of where big data alone was unable to answer important questions, a collaborative team working on behalf of NHS England analysed the electronic health records of more than 17 million adult NHS patients and linked them with 5,683 COVID-19 hospital deaths between Feb. 1 and April 25, 2020 to analyse their causes of death (The OpenSAFELY Collaborative, 2020). The researchers found strong associations between death from COVID-19 and being male, being old, being socially deprived, having diabetes, having severe asthma or other underlying health conditions, and being Asian or Black. They analysed 24 billion lines of structured data and their study had the advantages of high speed and scope. The study was able to detect some correlation between ethnicity and underlying health conditions, but not enough to explain the whole of the ethnic risk, nor could it say why this association existed. The study was useful, but it took a causal analysis to tease out the complex ways in which ethnicity did affect both the likelihood of COVID-19 infection and the severity of illness once infected. These included the overrepresentation of ethnic minorities among frontline health workers, and a higher likelihood of living in overcrowded, multigenerational households. These featured as infection possibility and background risk factors in my model. In fact, ethnicity when entered without any other observations in the model did not increase the probability of severe COVID-19 at all, but it did when each age group was considered one by one.

The reason for this is that White people, who make up the majority of the UK's population, tend to live longer than the other ethnic groups shown here, and old age is the single most important risk factor for a severe COVID-19 outcome. In an earlier version of this model, when COVID-19 prevalence was much higher, non-Chinese Asians and Black people actually came out at lower risk when age was unobserved. This is an example of Simpson's Paradox, when an effect that appears in a population considered as a whole disappears or is reversed when the population is broken down into groups that are analysed one by one. In the United States, where COVID-19 prevalence is still much higher than in Britain, this effect is more pronounced (Mackenzie, 2020).

The hospital population studied, although large, was not representative of the population. People in hospital tend to be older and sicker to begin with, and the study's method was to take those who had died of COVID-19 and then link back to their medical records. Taking COVID-19 deaths and then analysing the characteristics of those patients is not at all the same thing as taking someone from the general population with those characteristics and concluding something about their COVID-19 risks. This is an example of where Bayesian networks come into their own by being able to reason causally both backwards and forwards.

On July 24, Public Health England launched a major anti-obesity drive and argued that:

“evidence suggests excess weight is associated with an increased risk of the following for COVID-19: a positive test, hospitalisation, advanced levels of treatment (including mechanical ventilation or admission to intensive or critical care) and death.” (Blackshaw, et al., 2020)

The report synthesised a large amount of information and was useful for my project, but it was flawed in many ways. Firstly, 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, it 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 the other way – 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. However, clearing that observation and entering the observations “obese” and “morbidly obese” in the “obesity” node has no perceptible effect on COVID status measured to two decimal places. Drawing a causal diagram helped me understand and model why this was so.

## LIMITATIONS AND FURTHER WORK

The model, however, is only as good as the available data. Many revisions were made along the way as more was learned, for example about how various underlying health conditions affected vulnerability to more severe COVID-19 outcomes. In some cases, these were identified through association; in others, causal links were posited. A host of new symptoms was identified (this section was worked on by my colleague Rachel Butcher), leading to a downward revision of the proportion of asymptomatic cases. (This may have to be revised again as the effects of schools reopening start to be felt.) Many judgements had to be made on the basis of data from hospital patients, those who had been tested because of symptoms or those who volunteered for studies - all groups with inherent bias. The Office for National Statistics launched an infection survey pilot in late April to estimate COVID-19 infection in the community, which helped (Office for National Statistics, 2020). It is based on volunteers from an initial pool of 20,000 households who agreed to be tested. This provided the best estimates I could find so far of infection in the community but is still from a self-selecting group. (Including data from hospital patients and care-home residents did not perceptibly change the overall infection prevalence when weighted for proportion of the population.)

Data aside, the way the model itself is built has some weaknesses. Even though discretisation of continuous data is a known challenge for Bayesian networks, the binary categorisation of underlying health conditions as either high- or medium-risk is too crude. Observing more than one condition of the same level of risk does not increase the overall underlying health-condition risk, as it is currently constructed. In fact, at the last moment I discovered an inconsistency in that I had determined too rigidly which category someone would fall into. For example, someone of a healthy weight who had never smoked had been predicted to have a low risk of underlying medical condition. This made it contradictory to enter an observation of a medical condition with a medium or high risk. I adjusted the node probability table for “underlying medical condition” to allow for a little “wiggle room”. This was not an ideal solution and, given more time, I would have found something better.

Social deprivation/household income would have been a useful node to have. More links could have been drawn, for example from ethnicity to obesity or from religion to interactions with other people. These would have led to more nuanced results. The ethnicity node itself could be more differentiated. I did separate “Chinese” from “other Asian” because these ethnic groups have very different health traits. It would also have been useful to break down Black into, for example, African and Caribbean – again, because of different health traits and cultures. Pre-symptomatic could usefully have been added to the COVID status nodes, as there is some evidence that the viral load may differ between pre-symptomatic and asymptomatic people. These are all aspects that I would have developed, had I had more time.

## THEORY AND PRACTICE

The current COVID-19 pandemic provides a classic example of the type of problem that can be well-handled by Bayesian networks. It contains great uncertainty, incomplete data and fast-developing expert views about causality. I have done my best in the time available to untangle associations that have been observed between risk factors and COVID-19 outcomes and to draw causal relationships based on the best expert opinion available. For example, ethnicity is not directly related to COVID-19 status but is mediated by age, religion, occupation, underlying health conditions and other factors too. I have not committed the so-called prosecutors' fallacy of equating the probability of  $x$  given  $y$  to that of  $y$  given  $x$  – as described in the obesity example above. The model also exhibits the "explaining away" phenomenon that is characteristic of Bayesian networks. For example, entering the observation "severe" in "eventual COVID-19 status" causes the probability of being aged 75 to more than double from the prior probability. But then adding the observed probability "high" in "underlying medical condition" reduces the probability of being 75 or over, because there is an alternative explanation – and vice versa.

With more time, this model could explicitly include another characteristic strength of Bayesian networks – interventions. It could be extended to model the effects of mask-wearing, going back to school - or indeed using a digital contact-tracing app! If repurposed as an aid to policy-making rather than a personalised risk-assessment tool, it could also model interventions such as closing international borders or shutting down pubs, as well as competing utilities such as the health risks of reopening schools versus the educational risks of keeping children out of classrooms.

## LEGAL, SOCIAL ETHICAL ISSUES AND SUSTAINABILITY

The main ethical and legal consideration of this project, which is discussed in some detail in my dissertation, is privacy. Individuals need to feel confident that the information they enter about their health – some of the most private data there are - will be kept confidential and only used for the advertised purpose of personal risk assessment. The way the app is conceived is precisely to preserve privacy by only storing information locally and not using Bluetooth or any other technology to detect other users in the vicinity. In this sense, it is not a contact-tracing app like those using Apple and Google technology that have been adopted in much of the world, but far more of a risk-assessment tool. It has been conceived in such a way that the only data that would be centrally gathered would be the probability that the user has COVID-19, and the GPS location – with no personal data collected (Fenton, et al., 2020). This would be to allow epidemiologists to predict where local outbreaks may occur. The app also has the advantage that the user can enter more or less personal data at his or her discretion – a characteristic feature of Bayesian networks. There are no obvious sustainability issues with this project.

## REFERENCES

- Blackshaw, J., Feeley, A., Mabbs, L., Niblett, P., Atherton, E., Elsom, R., . . . Other members of PHE. (2020). *Excess Weight and COVID-19*. Public Health England. London: Crown. Retrieved July 28, 2020, from [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/903770/PHE\\_insight\\_Excess\\_weight\\_and\\_COVID-19.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903770/PHE_insight_Excess_weight_and_COVID-19.pdf)
- Cai, H. (2020, March 11). Sex difference and smoking predisposition in patients with COVID-19. *The Lancet Respiratory Medicine*, 8(4). doi:[https://doi.org/10.1016/S2213-2600\(20\)30117-X](https://doi.org/10.1016/S2213-2600(20)30117-X)
- Fenton, N. E., McLachlan, S., Lucas, P., Dube, K., Hitman, G., Osman, M., . . . Neil, M. (2020, July 19). A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing. London: medRxiv. doi:<https://doi.org/10.1101/2020.07.15.20154286>
- Jensen, F. V. (2009, December 1). Bayesian networks. *WIREs Computational Statistics*, 1(3), 307-315. doi:<https://doi.org/10.1002/wics.48>
- Mackenzie, D. (2020, July 6). Race, COVID Mortality, and Simpson's Paradox. *Causal Analysis in Theory and Practice*. Los Angeles, California: UCLA. Retrieved August 29, 2020, from <http://causality.cs.ucla.edu/blog/index.php/2020/07/06/race-covid-mortality-and-simpsons-paradox-by-dana-mackenzie/>
- Miyara, M., Tubach, F., Pourcher, V., Morelot-Panzini, C., Pernet, J., Haroche, J., . . . Amoura, Z. (2020). *Low rate of daily smokers in patients with symptomatic COVID-19*. Paris: medRxiv. doi:<https://doi.org/10.1101/2020.06.10.20127514>
- Office for National Statistics. (2020, August 28). *Coronavirus Infection Survey pilot*. Retrieved August 29, 2020, from [ons.gov.uk: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales28august2020](https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales28august2020)
- Pearl, J., & MacKenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.
- The OpenSAFELY Collaborative. (2020, July 1). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. doi:<https://doi.org/10.1038/s41586-020-2521-4>
- Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., & Imai, N. (2020, June 1). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet*, 20(6), 669-677. doi:[https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7)