Pandemic mortality
Covid-19 cases can be shaped by medical conditions and demography

Our risk model estimates chances of death and hospitalisation based on age, sex and comorbidities

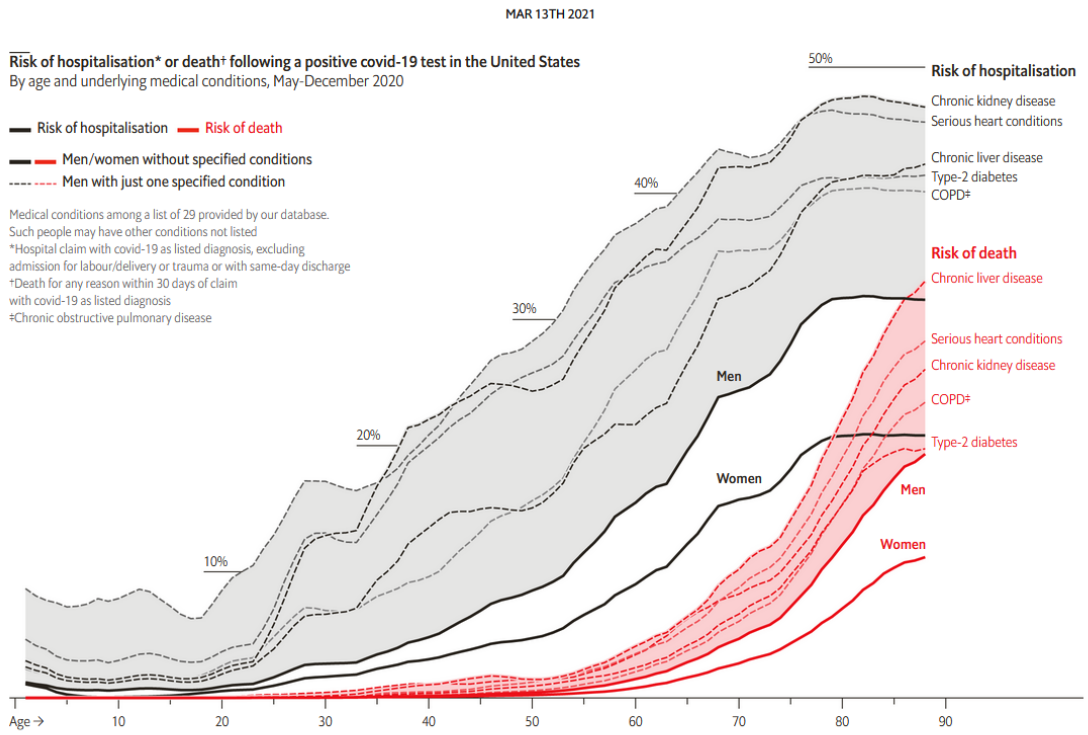
Graphic detail

The Economist, Mar 11th 2021

Covid-19 threatens everyone, but its risk is concentrated among particular groups of people. To help readers understand how the disease interacts with demography and with other illnesses ("comorbidities"), we have built a statistical risk model, using records in the Covid-19 Research Database from 425,000 people in America who tested positive. For any group of unvaccinated people of a given age, sex and mix of comorbidities, our model estimates the share that would be hospitalised or die within 30 days of a covid-19 diagnosis. To learn more about which medical conditions most exacerbate covid-19, please see Graphic Detail; the model's methodology is summarised here.

The interactive below lets you explore the model's output for any combination of variables. It assumes that comorbidities not selected are not present, even if they often appear together. For example, if you enter only Type 2 diabetes, you will receive an estimate for people with Type 2 diabetes but not hypertension. We do not store any records of which readers use the interactive, or of which medical conditions they select.

Users should not interpret these results as a personalised risk assessment. Any given individual's risk will differ from the group average that our model estimates. Readers seeking medical advice should consult a doctor. Rare combinations of inputs can produce unreliable results (see FAQ).



You can explore our interactive covid-19 risk estimator here, and read the methodology here.

EVER SINCE the covid-19 pandemic began, calibrating disease-related risk has

EVER SINCE the covid-19 pandemic began, calibrating disease-related risk has become a fixture of everyday life. Few places are fully safe. Is it wise to go to the supermarket? What about taking a taxi with the windows down while wearing a mask?

The reliability of our model's output depends partly on how many examples of the specified combination of age, sex and comorbidities are present in its training data. The more common a profile is, the narrower the confidence interval surrounding the central estimate. Enter an age and sex above, and you can see how the risk level for people with those attributes and the listed comorbidities compares with that of a representative sample of 10,000 people in our database.

Frequently asked questions

What information does the covid-19 risk estimator provide?

Our interactive tool estimates the risks posed by covid-19 to different groups of people. Formally, it is designed to answer a very specific question: if a group of randomly selected

people in the United States with the specified age, sex and comorbidities had been diagnosed with covid-19 on December 1st 2020, what percentage would have died or been hospitalised by the end of the year?

Why did you build it?

The broad contours of covid-19 risk are well-known: older and sicker people face more danger than younger and healthier ones do, and men more than women. However, there is far less awareness of the magnitude of these effects, how they interact with each other and which comorbidities are most relevant—particularly when assessing the chances of hospitalisation rather than of death. We built this estimator so that readers could explore a wide range of combinations of these variables.

What data is it based on?

The model is trained on medical records from the Covid-19 Research Database, drawn from 425,000 people in America who tested positive for the disease between May and December 2020. The archive lists their age, sex, date of diagnosis, presence or absence of 29 different comorbidities, whether they were hospitalised during their infection and whether they died in 2020.

The database has several limitations. It only includes people with health insurance, and does not list patients' location, ethnicity or date of death. Most octogenarians' ages appear as "80+". Anyone who has filed a medical claim since 2014 citing a comorbidity is listed with that condition, regardless of recency or severity—preventing distinctions between malignant tumours and cancers in remission. Not everyone without known illnesses is healthy: some have ailments not on the 29-condition menu.

Moreover, the database is not a representative sample of the SARS-CoV-2-positive population. Because it only contains records from people who have interacted with a medical service provider, it excludes those who weather the disease at home without medical assistance. As a result, the people it tracks are disproportionately old and sick. We have tried to counteract this bias using official data from America's Centres for Disease Control and Prevention, which record deaths and confirmed cases by age, sex and time period. However, our method could conceivably lead to a slight underestimation of the impact of comorbidities on risk.

How does it work?

The interactions between covid-19, age, sex and over 500m potential combinations of comorbidities—roughly 30,000 of which are present in the model's training data—are too complex to be captured by standard statistical tools such as logistic regression or proportional-hazards models. As a result, we used a popular machine-learning algorithm called "gradient-boosted trees", which is designed to incorporate such multi-faceted

relationships into its predictions. Please read our methodology summary for a detailed account of this process.

How accurate is it?

Our estimator is extremely reliable within the confines of its training dataset. To measure its accuracy on unseen data, we randomly split up the archive into two halves; trained separate models on each half; and used the resulting models to make predictions on the "opposite" halves. The estimator performed admirably: around 5% of people to whom it assigned a death risk between 4% and 6% did in fact die; roughly 30% of people to whom it assigned a hospitalisation risk between 29% and 31% were admitted to hospitals; and so on. When evaluating models' accuracy when making predictions on data not used to train them, the gradient-boosted trees fared far better than did simpler alternatives, such as logistic regression.

Whether such performance can be replicated outside the confines of the Covid-19 Research Database, however, is a different question. The circumstances of the vast majority of the world's SARS-CoV-2-positive population are not similar to those of the people in the model's training data. Barring the invention of time travel, no one in the future will be in the United States in December 2020, which is one of the model's central assumptions. More practically, most people who get covid-19 are unlikely to be infected with the same variants of the virus, have similar genetics or receive healthcare treatment of the average American quality at that time. We cannot estimate how any of these differences will affect our model's performance, but at least some of them are likely to cause significant errors.

Why does estimated risk sometimes decline at higher ages or with more comorbidities?

Such counterintuitive estimates result from quirks in our dataset. Many medical conditions tend to show up in pairs or trios. For example, almost 99% of people in our archive listed with hyperlipidemia also have metabolic disorders. Others are closely correlated to sex or age: 97% of people listed with breast cancer are female, and 96% of those with Parkinson's disease are 50 or older. People who do not fit these patterns—such as people diagnosed with hyperlipidemia but not metabolic disorders, men with breast cancer or young people with Parkinson's—are likely to have unusual presentations and perhaps particularly severe cases of these conditions, which may increase their vulnerability to covid-19.

Conversely, 99.9944% of the 500m potential combinations of comorbidities never showed up in our training data. Although the model can make educated guesses about such cases based on similar examples, its estimates will be quite unreliable. Moreover, unlike regression-based approaches—which treat all variables as having constant, independent effects that can simply be added together—gradient-boosted trees do not impose such assumptions. As a result, they are just as likely to predict that adding a comorbidity or raising a person's age will decrease risk rather than increasing it, in the absence of evidence to the contrary.

In particular cases, the model may well have good reasons for displaying surprising results. But in general, whenever it seems to behave strangely, the most likely explanation is that your instincts are right and that our tool, lacking sufficient data to produce a robust estimate, is wrong. You can tell how much to trust its predictions by the width of its confidence interval. The wider the shaded area surrounding the central estimate, the greater the chances that it is leading you astray.

What did you learn about covid-19 while developing this tool?

We were most surprised by the difference in the relative importance of age and comorbidities when measuring the chances of death versus those of hospitalisation. Whereas survival rates are primarily a function of age, even young people can easily wind up in hospital if they have medical conditions that sharply exacerbate covid-19's severity. We were also surprised to see that kidney and liver ailments, as well as cardiovascular ones, seemed to have more dangerous interactions with covid-19 than respiratory conditions did. You can read more about these findings in our Graphic Detail article.

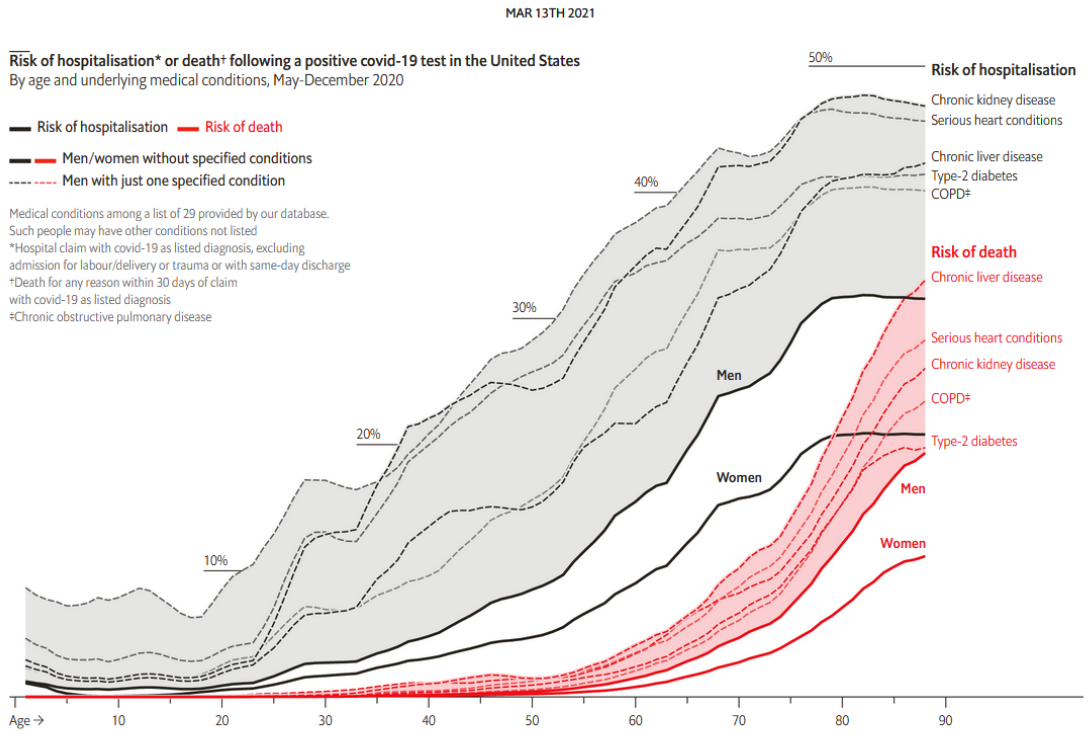
How should this information affect my personal risk choices?

It shouldn't—at least not without consulting a medical professional first. Leaving aside broad questions about how reliably the statistical relationships within the Covid-19 Research Database will translate to the world at large, our estimator doesn't have nearly enough information about individuals to provide accurate personal risk assessments. Its output represents rough group averages, from which every specific person's situation will differ substantially. Moreover, covid-19 can still harm people who survive it without a hospital stay. Many people suffer debilitating symptoms that persist for months, and even those who do not know they are sick can infect vulnerable people around them. Nonetheless, we hope that our estimator can cast some light on decisions that people will have to face by force of circumstance as societies open up. Once you know the impact of age, sex and the particular set of medical conditions that our model incorporates, you will have an inkling of the context in which other risks present themselves. ■

One size fits few
Our covid-19 model estimates odds of hospitalisation and death

Death rates depend mostly on age, whereas comorbidities sharply raise chances of hospitalisation in young people

The Economist, Mar 13th 2021 edition



You can explore our interactive covid-19 risk estimator here, and read the methodology here.

EVER SINCE the covid-19 pandemic began, calibrating disease-related risk has become a fixture of everyday life. Few places are fully safe. Is it wise to go to the supermarket? What about taking a taxi with the windows down while wearing a mask?

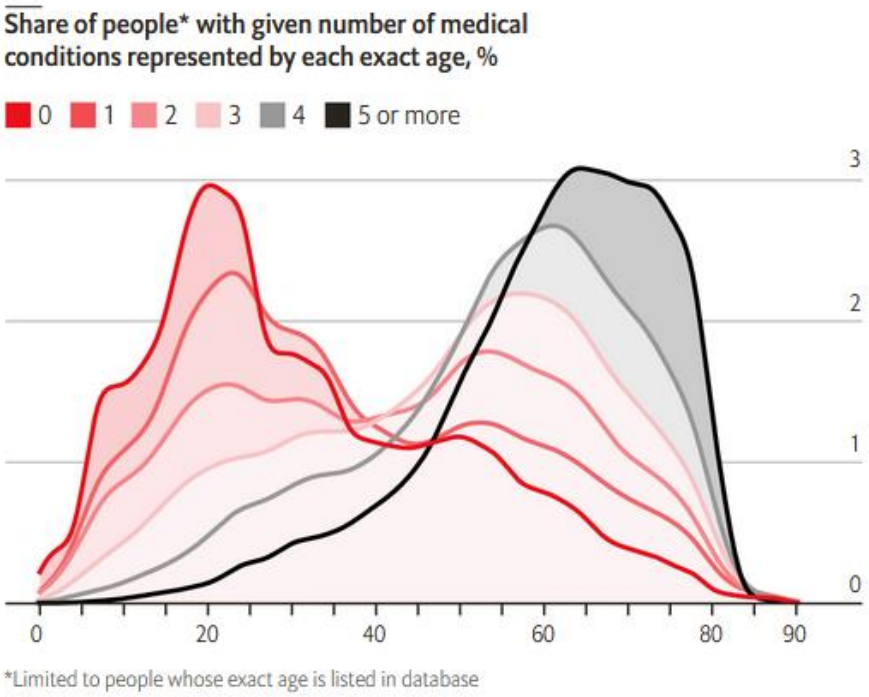
The answers depend not just on how likely an activity is to cause transmission, but also on how bad a bout with covid-19 would be for the individual involved. In rich countries the case-fatality rate (CFR) for people who test positive is just under 2% (the true death rate, including undiagnosed cases, is lower). Yet covid-19's lethality varies so much that most people do not face a low-single-digit CFR. Few children show symptoms, whereas the elderly—especially those with other illnesses ("comorbidities")—die at alarming rates. Officials have emphasised universal recommendations like masks and social distancing, leaving individuals to choose risk tolerances within those guidelines.

Such assessments can be complex. Although older people account for most deaths from covid-19, the mechanism behind this pattern is unclear. Are the elderly at risk purely because of their age? Or is it instead because they often have comorbidities that weaken defences against covid-19—and if so, which ones? There is no consensus about the relative importance of these factors. In America the list of comorbidities that enable younger people to get vaccines varies widely between states.

Making granular estimates of covid-19's risks requires lots of data. The sample needs to have plenty of rare examples, such as gravely ill teenagers and sprightly 90-year-olds. It also needs accurate proportions of specific demography-comorbidity pairings, such as men in their 30s with covid-19, pancreatitis *and* asthma.

Such a dataset now exists, though it has notable flaws. A group of American hospitals, doctors, insurers, pharmacies and data vendors have pooled data about their patients to create the Covid-19 Research Database, an archive of over 5bn medical records. In partnership with A3.AI, a research group that has spliced each patient's records together, the project's administrators have granted access to *The Economist*.

The archive records the age, sex and presence of 29 comorbidities among 104m people in America, of whom 466,000 were diagnosed with covid-19 in May-December 2020. It also lists which ones died in 2020, and, for people who tested positive, their date of diagnosis and whether they were hospitalised during their illness. Although the population in the dataset is sicker than average, this bias can be offset by adjusting the sample using official data on cases and deaths by age, sex and time period.

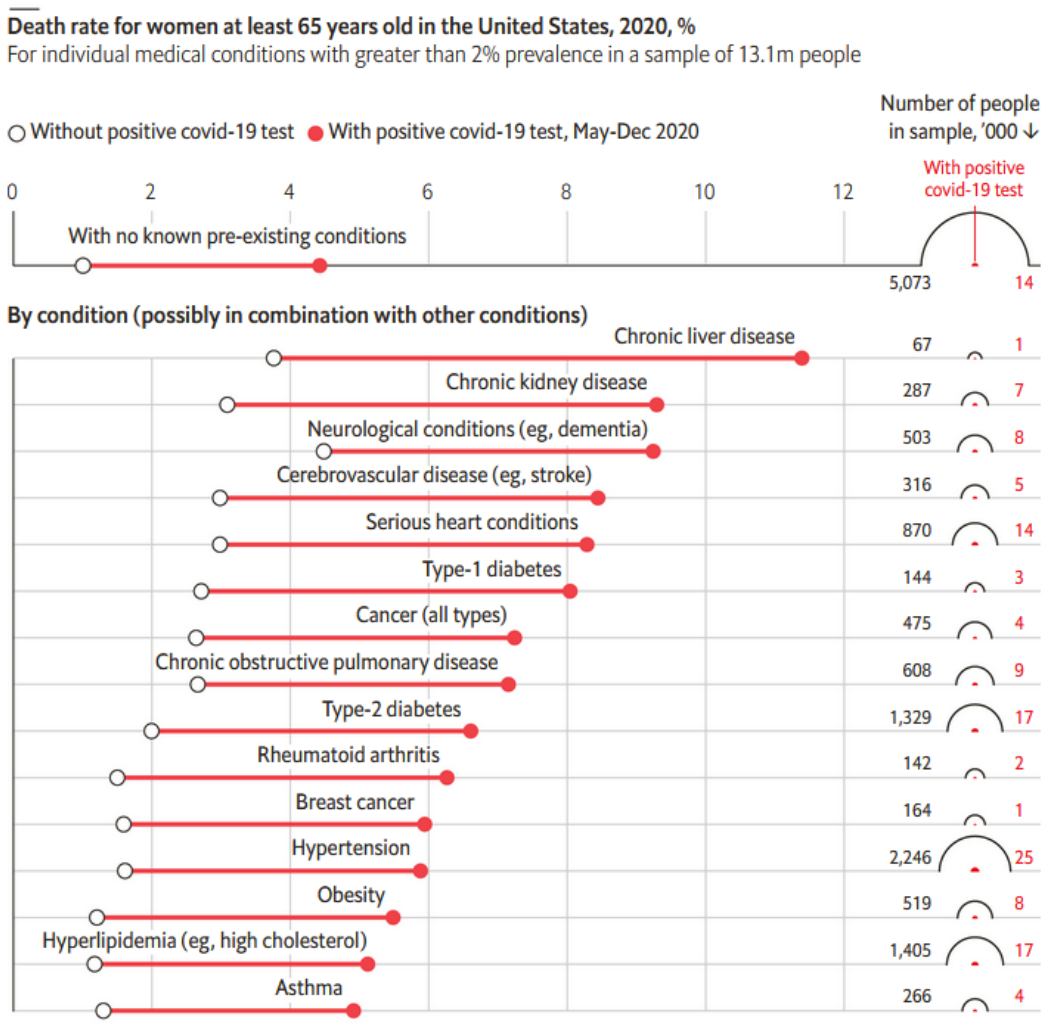


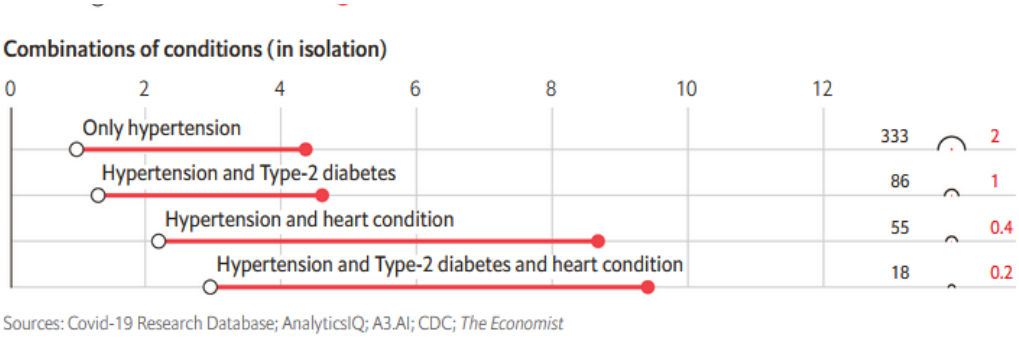
The data illuminate patterns that doctors have already seen in coronavirus wards, but are not yet conventional wisdom. Covid-19 spreads mostly through the air, and is often considered a respiratory ailment. However, complications in severe cases are often cardiovascular, including heart inflammation and irregular blood clotting. The archive bolsters growing evidence that covid-19 attacks the body broadly, and is most exacerbated by comorbidities that cause inflammation or that affect the circulatory system, such as kidney, liver or heart problems. In contrast, respiratory conditions like asthma matter less—though severe ones, such as lung cancer or pulmonary fibrosis, are big risk factors too.

How much danger such conditions pose to covid-19 patients depends on the outcome you measure. Most risk analysis focuses on deaths. On this metric, raw age is a stronger predictor than listed comorbidities; sex is important as well. After correcting for biases in the data, 8.5% of men and 4.9% of women in their 70s with no known conditions besides covid-19 died. The figures for people aged 25-34 with covid-19, heart disease and hyperlipidemia (eg, high cholesterol) were 0.8% for men and 0.7% for women. This means that you might want to wait until your tennis-playing, adventure-travelling grandparents are vaccinated before visiting them. It also means that governments have been right to vaccinate older people first, even unusually fit ones—and that men might need lower age cut-offs than women do.

However, covid-19 can cause people great harm even if it does not kill them. And when it comes to predicting hospital stays, comorbidities play a greater role. Of the 25- to 34-year-olds with heart trouble, hyperlipidemia and covid-19, a quarter of men and a fifth of women

were hospitalised—roughly the same shares as those of people in their 70s without other listed conditions. People aged 25-34 without known illnesses besides covid-19, in contrast, had just 1.6% (for men) and 1.0% (for women) chances of hospitalisation. Although most younger patients beat covid-19 eventually, those with relevant comorbidities often cannot do so at home.





Covid-19's interactions with demography and comorbidities are too complex for simple rules of thumb. To calculate risks for all possible combinations of these factors, we have built a statistical model using a machine-learning algorithm called "gradient-boosted trees". For any group of un-vaccinated people of a given age, sex and mix of comorbidities, the model estimates the shares that, within 30 days of a positive test for covid-19 in America in late 2020, would have died or been hospitalised. You can explore its output here and learn how we built it here.

Our model cannot estimate risk reliably for individuals. The archive used to build it has several limitations. It only includes people with health insurance, and does not list patients' location, ethnicity or date of death. Most octogenarians' ages appear as "80+". Anyone who has filed a medical claim since 2014 citing a comorbidity is listed with that condition, regardless of its recency or severity—preventing distinctions between cancers in remission and malignant tumours. Not everyone without known illnesses is healthy: some have ailments not on the 29-condition menu. And our correction for the data's overrepresentation of sick people could yield an underestimate of the impact of comorbidities.

Moreover, the model's assumptions may not hold up beyond its training data. It assumes that people are infected with one of the strains of SARS-CoV-2 common in America in late 2020; that their quality of treatment and odds of getting tested are similar to the American averages at that time; and that they demographically and immunologically resemble people in the data who share their listed attributes. In fact, new viral variants are spreading fast; most health care is either better or worse than the American average; treatments keep improving; and differences in genetics and past viral exposure can affect CFRs.

Yet despite these limitations, the model is probably more accurate than most publicly available alternatives. It accounts for interactions between comorbidities rather than taking each one in isolation, and estimates chances of both hospitalisation and death. Its output represents a group average, from which any given individual will differ. But by narrowing that group to people of the same age, sex and mix of comorbidities, it provides a more relevant starting point than a one-size-fits all average.■

Sources: Covid-19 Research Database; AnalyticsIQ; A3.AI; CDC; *The Economist*