Pandemic modeling, good and bad

Abstract
What kind of epidemiological modeling works well, and what kind doesn't? This is determined by the nature of the target: the relevant causal relations are unstable across contexts, which tells against any modeling that assumes otherwise. I look at two influential examples from the Covid pandemic. The first example is the paper from Imperial College, London, that projected future infection rates under various policy scenarios, and that in March 2020 was influential in persuading the UK government to impose a lockdown (Ferguson et al 2020). Because it assumes stability, this first example of modeling fails: it carries no epistemic force at all. A different modeling strategy is required, one that is less ambitious but more effective. This is illustrated by the second example: the paper, also from Imperial College, London, that in December 2020 first estimated the transmissibility of the Alpha variant (Volz et al 2020). This second, contextual example of modeling works well.

Keywords: pandemic, model, epidemiology, Ferguson, fragility

1. Introduction

What kind of epidemiological modeling works well, and what kind doesn't? There has been a vast amount of epidemiological work in response to the Covid pandemic. I look at two high-profile examples, each of which greatly influenced policy, in the UK and elsewhere. The first example is the CovidSim model in a paper by Neil Ferguson and colleagues at Imperial College, London (Ferguson et al 2020), the dire projections of which were a large factor in persuading the UK government to impose a national lockdown in March 2020 (a reversal of its previous policy). The second example is a paper from December 2020 by Erik Volz and colleagues (Volz et al 2020), also at Imperial College, London, which was the first to establish, and estimate, the greater transmissibility of the Alpha variant.

The lesson of these examples is that taking a model to apply stably to many contexts fails. Epidemiology is not like Newtonian physics, where a single master model can be developed and applied. The reason is that its target causal relations are unstable. Unlike Newtonian laws, they do not hold reliably across cases, but instead hold only intermittently and unpredictably. Extensive case-specific investigation is needed each time to know which – if any – model applies, and so a contextual rather than master-model strategy is required. As we will see, this is why the models in (Volz et al 2020) succeed but the CovidSim model fails.

Methodologically speaking, this distinction between contextual and master-model strategies is central. It cross-cuts previous distinctions in the epidemiological literature, such as that between model-types (agent-based, compartmental, curve-fitting), and that between interpreting models causally or non-causally (Fuller 2021). Each model-type, and both causal and non-causal models, can be developed and used in accordance with either a contextual or master-model strategy. It is the latter distinction that matters.

'Context matters' is a familiar refrain across social sciences. In this paper, I connect it to an underlying contrast between stable and unstable relations, in order to explain when and why some modeling strategies work better than others. In Section 2, I set out the background philosophy of science, using a simplified, non-pandemic example. In Section 3, I discuss the CovidSim model. In Section 4, I discuss the Volz paper, and informal methods.

2. Philosophical background: two strategies

Imagine two systems. In one, causal relations are stable and reliable; in the other, they are intermittent, often fleeting, and unpredictable. In this section, I explain why optimal methodology depends on which kind of system we are investigating (Northcott forthcoming). I illustrate with a simplified, non-pandemic example.

Suppose we want to predict the motion of a newly discovered moon. We can apply a Newtonian two-body gravitational model, inputting the moon and parent planet's masses and current motions. Something like this procedure is a staple of actual space exploration. Why does it work? The key reason is *stability*: the Newtonian model that has been successful elsewhere can be assumed still to apply, because gravity itself can be assumed still to be operating in the same way. In each new case, just re-apply the same Newtonian master model. Call this the *Master-Model* strategy.

Master-Model works well even in the face of *noise*, which is when there are significant influences not captured by our model. For example, the moon's motion may be deflected by gravity from a second moon, by impact with a comet, or (at least for a small moon maybe) by human interference, in which case, because of these disturbing factors, the Newtonian master model no longer predicts accurately. Nevertheless, the model still reliably identifies *one* of the factors influencing the moon's motion, namely the gravitational interaction between moon and planet. In this sense, the model still explains 'partially' (Northcott 2013). To explain fully, or predict accurately, we must add in the effect of unmodeled disturbing factors. This strategy – of developing a master model and then in specific cases adding in disturbing factors as needed – was already advocated by Mill almost two centuries ago (1843). It has been a staple of philosophy of science about modeling, as many authors have focused on how models, even if idealized and even in the face of noise, may nevertheless succeed by isolating stable causal tendencies or arrangements (Cartwright 1989, Mäki 1992).

In this way, a master model provides partial understanding even in the many cases when empirical accuracy is imperfect. Such an achievement, on this view, is even *superior* to mere empirical accuracy. Why? Because empirical accuracy in any particular case requires taking account of every local factor, no matter how sui generis or transient. But what is of greater interest to science, as a pursuit of systematic knowledge, is those factors that generalize – which is just what a master model captures.

Master-Model relies on stability in two ways. First, stability is essential metaphysically. A master model is a reliable base onto which case-specific disturbing factors can be added, only because the relations it describes are stable. (Mill himself was well aware of this: he had in mind the practice of economics, where he thought core psychological tendencies such as seeking to increase one's own wealth are indeed stable in the required way.) Second, stability is also essential epistemologically. If we are lucky, warrant to apply a master model comes from empirical success here and now: the Newtonian gravity model, for example, is given warrant by successfully predicting the motion of the moon. But often there is no empirical success here and now, because of noise. Then, warrant can come only indirectly, from empirical success elsewhere: we have faith, for example, that even when noise means it predicts badly here and now, still the Newtonian model has correctly identified one gravitational force at work. Why? Because of the model's empirical success elsewhere.

But such indirect warrant is justified only when there is stability. It is only because gravity operates in the same way across cases that the Newtonian model's warrant from elsewhere stays good over here.

So, what Master-Model requires is stability; noise is irrelevant. *Without* noise, a master model is empirically accurate across many cases only when the relations it describes operate in a stable way. *With* noise, meanwhile, while a master model is no longer empirically accurate, now we may retreat to Mill's strategy, confident that a master model does at least accurately describe some of the factors at work, even if additional disturbing factors are present too. Master-Model is effective with or without noise – but only given stability.

But there is an alternative – for when stability is absent. To introduce it, imagine now a different moon example. This time, the 'moon' is a toy moon on a string, being carried by a child around a toy planet. How might we predict the motion of *this* moon? The best candidate here for a master model is probably something psychological, perhaps that a child will continue an action they are enjoying. Call this the Continue model. So, if the child has carried the toy moon around the planet for two 'orbits' happily, the Continue model predicts that the child will continue for another two orbits. This prediction will be right sometimes. Other times, it will not be: perhaps the child gets distracted, interrupted, or bored, or perhaps they are following instructions in an online science class (two orbits only), or perhaps they are playing a game with a friend (take it in turns to hold the moon). The underlying problem is lack of stability: the relations captured by the Continue model do not reliably apply across cases. Using just a single model is no longer effective.

What alternative works better? Many models are available, some of which we might think of as loose hypotheses or rules of thumb as much as formal models. Different models describe how a child behaves: with friends, in a school class, when they are tired or bored, when they are interacting with a sibling, when they are affected by poverty or divorce or moving to a new house, and so on. The key is which of these models applies in any particular case. To discover that requires much case-specific work, looking for contextual clues and triggers: the character of this child, the nature of this household, is the child tired late in the day, is the child hungry, is the child – or friend or parent – generally frustrated after a prolonged lockdown, is the weather bright and warm or is it gray and miserable, and so on. In short, exactly the things a parent actually considers when trying to understand a child's behavior. Instead of a single master model, we choose from many different models, case by case.

Label this new methodological strategy, *Contextual*. Contextual is not against cross-contextual models as such; indeed, the larger the available toolbox of such models, the better. Rather, Contextual implies two things. First, a change in balance: choosing between lots of alternative models rather than always going with one master model means that, overall, relatively more scientific effort must be devoted to local empirical investigation. Because no single model is assumed always to apply, we must be more sensitive to local detail, in order to select wisely from our toolbox. Second, a change in how models are developed. They should not be developed in the abstract, relying on real-world stability to keep them applicable. Instead, models must be developed by empirical refinement, learning by continually applying them to real-world cases (Ylikoski 2019). This might sound obvious, but it is often not done. We will see an example in the next section.

Without stability, as noted earlier, the warrant of predictive success cannot be imported from elsewhere. That means accurate prediction is always required here and now. But without stability, accurate prediction is harder: the toy moon's motion is harder to predict than is the real moon's. It is not always easy to know which model applies, and noise is ubiquitous. But that, as it were, is nature's fault, not ours. Still, this is not a counsel of despair: we can get a decent grip on the toy moon's motion sometimes, some predictions are more accurate than others, and some explanations are fuller and better warranted. It is up to us to find them. In toy-moon circumstances, Contextual is the best way to do that.

In sum, what matters methodologically is stability. When the target relations are stable, it is best to investigate via a single master model (assuming an accurate one can be found), such as a Newtonian model of gravity. This strategy is effective even when empirical accuracy is disrupted by noise. But when target relations are unstable, local investigation is required each time to discover which of many candidate models might apply, and any model selected needs to be empirically accurate there and then (Northcott 2017).

The core contrast, on this paper's account, is between stable and unstable relations. This is not the same as the contrast between model monism and model pluralism (Teller 2001, Veit 2019, Cartwright 2019). It is true that Master-Model is a form of monism and Contextual will often lead to pluralism. But in principle, one could have many different stability-based models, each capturing different stable relations that might be present. Conversely, one response to instability could be to stick monistically to a single context and single model. Further, Contextual gives its own, independent rationale for pluralism, namely that unstable relations may demand different models as those relations change. Contextual also informs the details of how model pluralism should be implemented.

Armed with this philosophical background, turn now to the Covid pandemic. Which methodological strategy to choose? I answer: Contextual, not Master-Model. The target relations are unstable. As it were, the pandemic is more akin to toy-moon than to real-moon.

3. Pandemic modeling: Against the Master-Model strategy

3.1. The CovidSim model

March 2020 was a key moment in the UK's pandemic response. Infections were rising rapidly, and decisions had to be made in a hurry: how many would die given various policy interventions, or given no interventions at all? Several competing models sought to advise. An especially influential one was developed by Neil Ferguson and colleagues at Imperial College, London (Ferguson et al 2020). It was one of the main factors that persuaded the UK government to reverse its previous policy and to impose a national lockdown (Grey and MacAskill 2020, Ford 2020, Conn et al 2020).

(Ferguson et al 2020) does not itself develop a new model; rather, it adapts a transmission model originally developed by the same team for influenza outbreaks. This Covid model – CovidSim – does not simply analyze the dynamics of population-level aggregates, in the manner of the classic SIR (Susceptible-Infectious-Recovered) model. It is more ambitious. It models interactions between individuals within the home, at school, at work, and in the community. (The model is a hybrid of agent-based and compartmental.) Various parameter values are estimated from actual data, such as age and household size, average class sizes and staff-student ratios, and workplace size and commuting distances, all referenced to population density data from the census. Covid-specific values are added for the virus's incubation period, reproduction number (i.e., its 'R-number'), and infection fatality rate.

Using CovidSim, (Ferguson et al 2020) analyzes two kinds of policy intervention. The first it labels Suppression: use restrictive measures to drive the virus's R-number below 1 and so reduce case numbers, in the hope that vaccines or treatments with high efficacy are eventually developed. Until vaccines or treatments arrive, restrictive measures must be re-imposed periodically because infections will rise again whenever measures are eased. The paper labels the second kind of intervention Mitigation: less strong restrictive measures, which seek to reduce the R-number but not all the way to below 1. Rather, the aim is to spread out the number of cases over time (compared to no intervention) so that at no point are health services overwhelmed, in the hope that immunity in those previously infected builds up gradually in the population, leading eventually to a decline in case numbers. This eventual decline would happen even without vaccines. The paper concludes, based on its simulations, that any feasible version of Mitigation would leave national Intensive Care Unit (ICU) capacity overwhelmed (by a factor of at least 10) and would inevitably lead to a very large number of deaths. Therefore, Suppression is preferred. Acceptance of this recommendation by the UK government was the paper's major policy impact.

In which methodological camp does CovidSim belong, Master-Model or Contextual? On one hand, it aims to be more sensitive to local variation than are traditional epidemiological models such as SIR, and accordingly it has many more parameters. But overall, it clearly belongs in the Master-Model category. It is intended to apply to many infectious-disease epidemics. In each application, as with the Newtonian model of gravity, predictions are generated by applying the same underlying model with the same structural relations and parameters; the only variation is in the values of those parameters.

I will be critical below. But I want to acknowledge that the CovidSim model – an intricate piece of work – was produced, at speed, at a moment of national crisis, and was a sincere, well-intended attempt to guide policy.

The critiques below are structured. The first, and most important, critique is the lack of empirical confirmation (Section 3.2). I then consider a fallback defense, namely that CovidSim can gain warrant even without empirical confirmation. I reject this fallback defense on the grounds, first, that relevant parameter estimates are inaccurate (Section 3.3), and second, that CovidSim omits important relations (Section 3.4). After thereby establishing that CovidSim lacks warrant, I diagnose why. The underlying problem is that it treats as stable relations that in fact are unstable (Section 3.5). This undercuts hope that the model can be augmented fruitfully in the future. I conclude that none of the benefits of Master-Model are realized in this case, and that CovidSim has no epistemic value (Section 3.6).

3.2. Lack of empirical confirmation

The most important difficulty is lack of empirical confirmation. As many have noted, assessing this needs care because the CovidSim model, in keeping with its policy-advisory role, gives only projections (Fuller 2021, Schroeder 2021). In other words, its predictions are conditional: how many infections, and how much pressure on health services, *would* there be given various policy combinations? Actual policy did not exactly replicate any of these combinations. Are criticisms of the model's poor predictive record therefore unfair?

In reply, two points. First, if a model really cannot be tested empirically, that is a bad thing, not a good thing. Confirmation will remain absent forever. The model will not receive empirical feedback or refinement, blocking off the main route to scientific progress.

Second, in fact, the CovidSim model can be tested, at least somewhat. For policy scenarios similar to what turned out to be the actual policy mix, were its projections approximately right? No, they were not. There has been some scholarly to and fro on this question (e.g., Winsberg et al 2020, 2021, van Basshuysen and White 2021a, 2021b), mainly with regard to infection numbers and ICU usage. Winsberg et al (2021) argues that the model greatly overestimated deaths and ICU usage in Sweden and Florida. Generally, other projections of CovidSim were also inaccurate, such as how quickly deaths would accumulate, how long policy measures would need to be maintained to avoid ICU capacity becoming overwhelmed, and how quickly policy measures would then need to be re-introduced to prevent case numbers rising too high again. A similar model also clearly applied poorly to places such as Africa, where infection numbers were far lower than predicted. Various anecdotal failures cement the same negative verdict. SAGE (the official scientific advisory body to the UK government), informed by CovidSim, commented on 20th March: 'It is very likely that we will see ICU capacity in London breached by the end of the month, even if additional measures are put in place today.' Additional measures were put in place, but London's ICU capacity was not breached. And Neil Ferguson himself in late March 2020, after lockdown had been imposed, and again informed by the CovidSim model, predicted to Parliament that total UK deaths would top out at about 20,000, which of course sadly proved a huge underestimate. And there is no track record of predictive success in other epidemics to confirm the model independently. (Such success elsewhere would only be relevant anyway if we assumed – implausibly – that the target is stable enough for warrant to carry over from, say, an influenza epidemic to the Covid one.)

Prediction in epidemiology is challenging generally (Broadbent 2013), so this poor performance is not surprising, and should not be jeered at. The point is only to note its epistemic implications.

Might the CovidSim model accurately capture some of the causal relations involved, and so still be partially explanatory? That is certainly possible, metaphysically speaking, but the point is not terribly helpful. What is needed is reason to take CovidSim to be probative, for public policy and as a projection tool. And that means warrant.

3.3. Unrealistic assumptions

CovidSim is idealized, in other words it makes many false assumptions. Does this matter? Not necessarily. A lot of recent philosophy of science has addressed this question, and the consensus is that, roughly speaking, idealized models can explain when their falsity does not matter, in other words when their idealizations are true enough for the purposes at hand. The sting in the tail, though, is that empirical accuracy is still required at some point. A Newtonian model of gravity, for example, makes false assumptions, yet because it predicts accurately, and because for many purposes its assumptions are 'approximately' true, it is widely (and rightly) considered explanatory. But the predictive endorsement is essential, and that is just what CovidSim lacks.

But there is an alternative defense. A model's assumptions might be so compelling that empirical confirmation of the model's results, while desirable, is not essential. Even if not confirmed empirically, on this view, a model retains some epistemic force as a guide for interventions, and it can explain partially. (Mill believed – mistakenly in my opinion – something similar of economic models built around the assumption that humans seek to maximize their wealth.) But for the CovidSim model, this defense is unconvincing, as we will see now.

A few of CovidSim's parameters can be estimated relatively straightforwardly. Already, difficulties arise. Two of these 'easier' parameters are the virus's R-number and its infection fatality rate. The model estimates these only to within an approximate range: 2.0 to 2.6 (R-number) and 0.25% to 1% (infection fatality rate). But subsequent work suggests that, in both cases, the true values may well have been outside these ranges. In particular, in a society such as the UK without social distancing measures, a central estimate of the R-number (i.e., R_0 for the original variant) is 3.0 (Billah et al 2020), while a central estimate of the infection fatality rate in the UK's first wave is 1.1% (Brazeau et al 2020), albeit new treatments lowered this number later. Further, CovidSim's estimates are meant to cover all of the policy scenarios that it analyzes, but they do not. The scenario of zero social policy interventions, for example, produced, for the UK case, the headline-catching projection of 510,000 deaths. Yet with zero policy interventions, ICU capacity (and hospitals generally) would likely have been swamped – by a factor of 30 according to CovidSim's own estimates – so that only a tiny proportion of seriously ill Covid patients could have received hospital treatment. In that circumstance, given that the hospitalization rate in the UK of Covid patients in the first wave was about 4 to 5%, the infection fatality rate would likely have been much above CovidSim's estimate of 1%.

And, to repeat, these difficulties are with regard to the R-number and the infection fatality rate, two of the easier parameters to estimate. Other parameters required a lot more work. Some could be estimated only by means of educated guesses – because the data for anything more than that didn't exist. If we require symptomatic Covid patients to stay at home, then how many days at home? It is estimated 7 days. What will be the effect on patients' contacts? It is estimated that non-household contacts will decline by 75%, while within-household contacts remain unchanged. How many households will comply with this policy? It is estimated that 70% will. Values for other parameters are educated guesses too: that a symptomatic patient's household members will comply with voluntary home quarantine 50% of the time; and that 75% of those over 70 will comply with social distancing, reducing their contact by 50% in workplaces and by 75% in the community, while increasing it by 25% within households. Educated guesses were made also about the impacts of social distancing of the entire population, and of schools and universities closing. How long these various measures last is also relevant. The model estimated that each would last three months, except for social distancing of those over 70, which would last four months.

Inevitably, some of these educated guesses turned out to be more accurate than others. Notably, many policy measures, such as two-meter physical distancing, mandatory mask-wearing in indoor public spaces, and closure of nightclubs, ended up in force in the UK for almost a year and a half – far longer than the three months assumed in the model.

These details cast serious doubt, to say the least, on whether CovidSim's assumptions are compelling. But perhaps its projections are not sensitive to the precise parameter estimates, so that all we need is for those estimates to be roughly correct? If so, then we need to know how roughly. (Ferguson et al 2020) does report one sensitivity analysis – but of very limited scope. This analysis shows that the main policy recommendation, namely in favor of the Suppression policy over the Mitigation one, is not sensitive to the precise values of the R-number, the infection fatality rate, or the number of cases that triggers policy interventions. Or at least, the analysis shows this for values of these parameters that the paper considers plausible, such as between 2.0 and 2.6 for the R-number (remember, likely the R-number's actual value was above this range). But there are many other parameters in the model: how sensitive is the model's main policy recommendation to those? And how sensitive are its detailed quantitative projections? Given the number of parameters in the model (over 900), it seems doubtful that a full sensitivity analysis for the main conclusions is even feasible. Certainly, none is given in the paper. (As with several shortcomings, Ferguson et al 2020 does acknowledge this problem, but offers no solution.)

Two projections of the CovidSim model do seem clearly to fail a sensitivity test. The first is the estimate of 510,000 UK deaths in the absence of any policy intervention. This projection assumes an infection fatality rate of about 1%. But, as noted above, if ICUs and hospitals are overwhelmed, the infection fatality rate would likely be much above 1%, increasing the total number of deaths in direct proportion. An infection fatality rate of 2%, for example, would raise the figure for deaths to 1 million. The second projection is that, even on the best Mitigation scenario, 250,000 would die in the UK. This projection is central to CovidSim's main policy advice. But note how sensitive it is to a single assumption, by the paper's own admission: 'In the UK, this conclusion has only been reached in the last few days, with the refinement of estimates of likely ICU demand due to COVID-19 based on experience in Italy and the UK (*previous planning estimates assumed half the demand now estimated*) and with the NHS providing increasing certainty around the limits of hospital surge capacity' (16, italics added). That is, the estimated number of deaths under Mitigation doubled almost overnight.

More generally, later work suggests that CovidSim's projections are highly sensitive not only to its estimates of parameter values, but also to omitted factors (see the next two sections), and to uncertainty about which conditions actually apply (Edeling et al 2020, Winsberg et al 2021). This reinforces the point that the model's own sensitivity analyses are not enough. True, time was of the essence, so there is a limit to how much sensitivity analysis was feasible. But this practical constraint does not alleviate the epistemic problem.

3.4. Omitted relations

Incorrect parameter estimates are only part of the problem. In addition, the CovidSim model also omits many relations. Consider just two mentioned by (Ferguson et al 2020) itself. First, degree of social contact is influenced by people's spontaneous behavioral responses: high case numbers tend to lead to more social distancing spontaneously, and conversely low numbers lead to less distancing. Second, if schools are closed, this reduces health service capacity because some health workers who are also parents are forced to stay at home. Each of these omitted relations implies that parameter values assumed by the model to be constant – namely, levels of social contact and health service capacity – are, in fact, functions of other variables in the model. CovidSim's assumed constant values for these parameters therefore clearly are not so compelling as to be self-warranting.

3.5. No jam tomorrow: instability

The above range of failures is true of rival models too. None is predictively endorsed, and all make unrealistic assumptions and omit relevant relations. Is there, nonetheless, still hope for Master-Model? Perhaps the pandemic is characterized by stability plus noise. If so, then even though no current master model has acquired empirical warrant, some future master model may do better. Perhaps any omitted relations could be added to a hypothetical grand super-model and then given due weight case by case.

But, alas, disappointment with Master-Model is likely to be permanent. Why? Because of instability. We have just noted one source of such instability: the omission by CovidSim of some relations renders other, un-omitted relations – i.e., relations that are in the CovidSim model – unstable. But there is a more fundamental problem. Many key relations (not just parameter values) are unstable inherently, not just because they have been mis-specified by the model because of omissions. These relations vary across time, place, and virus type; they do not hold generally. By their very nature, relations that vary across cases in this way are not easily captured by a single master model, so Master-Model will always be unsuitable. It would be as if the way that gravity works varied from case to case, and gravity applied at all only to some planets and moons but not others. That would leave our Newtonian two-body model inevitably inadequate in many cases, at best one modeling option among many.

Consider, say, border controls – a key policy tool for many countries. How do border controls impact on infection rates? The answer is awash with context-specific details, varying with the idiosyncrasies of a country's borders, trading flows, location on transit routes, number of residents with ties abroad, number of border personnel and hotels, not to mention the nature and evolution of the relevant virus or disease. When should border controls be triggered, for how long maintained, how long should quarantine periods be, and what exceptions allowed (some countries? only some freight? airline personnel?)? All of these factors varied between countries that implemented border controls, and varied over time.

Similar remarks apply to test and trace programs. Compare the programs needed for Covid with those needed for SARS in 2003, or, within the UK, compare outsourced private operations with those run through local councils' public health officers. Similar remarks apply to public health messaging too. Compare, say, the very different media and political environments in Singapore, UK, and USA.

None of border controls, testing programs, or public messaging were modeled by CovidSim. But many relations that were modeled by it are likely unstable too. Popular resistance to lockdown measures has varied greatly across countries, over time, and much more for some measures than for others; local factors explain the differences, and thereby impact on the key factor of how long a policy can be sustained. The value of the R-number is not a constant but rather is a function of environments, behaviors, and political decisions. So, too, is the time between infection and transmission. (Ferguson et al 2020) states (8) that 'stopping mass gatherings is predicted to have relatively little impact'; but while this may be true for influenza, community transmission of Covid seems to be powered disproportionately by large 'super-spreader' events, which implies that stopping mass gatherings does have a large impact after all. As a result, CovidSim missed the disproportionate number of deaths in nursing homes and other care facilities. The relevant relations are unstable across influenza and Covid epidemics. (More generally, CovidSim omits social network effects, and therefore misses potential targeting of interventions at particular hub actors (Manzo 2020).)

This underlying problem of instability is not a result of the Covid virus – unlike planetary orbits – being new to science: the problem will not go away with advancing knowledge. New knowledge certainly helps with Contextual modeling but, if anything, it causes the target relations here to become less, not more, stable. The efficacy of government restrictions, for example, may change as it is learnt how to implement them better, as vaccines become available, or as public obedience wanes.

For comparison, in the gravity example the moon's position and velocity vary continuously, but this is not a problem because a Newtonian model tells us not just the effects of that variation but also when to expect it. There is no 'surprise' variation in gravity's influence *that* *requires knowledge from beyond the model* to predict and understand. But the same is not true of the pandemic. For instance, the determinants of the R-number's value, and the precise relations between the R-number and other components of the CovidSim model, are each unstable across cases because they depend on contextual factors such as virus characteristics, local history, and local politics. We require knowledge from beyond the model itself to update the model's correct form.

3.6. Verdict: Master-model and efficiency

Master-Model's most important methodological virtue is *efficiency*: a master model is a short-cut to successful predictions, interventions, and explanations. But with CovidSim, this efficiency gain no longer exists. Too many target relations vary case by case, and so Master-Model is misconceived from the start. This is the fundamental problem.

Certainly, the pandemic prediction models so far have been nowhere near the efficient Master-Model ideal. Even if predictive success were achieved – which it hasn't been – it could only be via so much contextual work each time that it defeats the point of Master-Model in the first place. Indeed, for CovidSim, the required contextual work is so great that in practice it is doubtful it can be done. Corners were unavoidably cut. The large number of target relations that are unstable suggests that these difficulties are permanent. In the next section, I discuss how to follow the Contextual strategy instead.

Where does this leave the CovidSim model? In my view, its advice carries no epistemic weight at all. Its predictions have a track record of inaccuracy, and the model is not empirically confirmed in any other way. It omits many important relations, many of the relations it does include are likely mis-specified, and many parameter values are estimated inaccurately. No sensitivity analysis reassures us that these errors are not fatal. According to no philosophical theory of explanation does such a model explain (Northcott and Alexandrova 2013), which here means we have no warrant to think the model has captured the true causal structure, and so no warrant to trust it as a guide to interventions.

The literature on idealized models does offer two other potential defenses, but neither of them helps here. The first defense is that, even when lacking empirical warrant, idealized models may give 'how-possibly explanations': rigorous accounts of how things work in a hypothetical world in which the model's assumptions are true. The hope is that how-possibly explanations shed indirect light on the actual world, perhaps by illustrating how things could be or might have been (Grüne-Yanoff 2009, Aydinonat 2008, Forber 2010). But (Ferguson et al 2020) explicitly aims to model the actual world – to its credit, given that its goal is to advise government policy. A second defense of idealized models is that, even when they do not themselves predict or explain, still they might be useful heuristically, perhaps by directing our attention to important factors otherwise neglected or overlooked (Alexandrova 2008, Alexandrova and Northcott 2009). But CovidSim actively turns our attention *away* from key omitted factors. And it does not guide us towards the kind of work that, as we will see now, actually delivers.

In a crisis, speed matters. (Generally, a constraint on optimal methodology is the context in which a question is being investigated.) So, Master-Model does carry a pragmatic advantage: a model is available "off the shelf", with less need for time-consuming local investigations. And CovidSim was certainly adapted quickly from its influenza origins. But not only speed matters. If a model lacks epistemic force, that defect trumps speed.

The CovidSim model influenced policy when it shouldn't have. It might be that, by inspiring a lockdown, it influenced policy for the better, and so on this occasion it had highly beneficial consequences. If so, then that was by luck. Judgment that the new policy was indeed better should not be based on CovidSim, but rather must be based on accumulated experience from many countries, assessed by other methods.

4. Pandemic modeling: In favor of the Contextual strategy

Turn to a more fruitful path. Any attempt to establish causal relations must work with a 'model', i.e., with a posited causal structure, even if just a singular causal claim. The distinction between the Contextual and Master-Model strategies does not lie in whether they use models – both do that – but rather lies in the *way* that they use them. Contextual urges that model development be tied closely to empirical feedback, and that, in contrast to one-size-fits-all, choice of model be carefully tailored to particular target. This is best shown by example.

4.1. Transmissibility of the Alpha variant

The paper that first established the higher transmissibility of the Alpha variant was written in December 2020 by Erik Volz and colleagues at Imperial College, London and elsewhere. Unlike those of (Ferguson et al 2020), this paper's conclusions are compelling.[1] Even though they apply directly only to one instance, because they are established reliably they serve as building blocks for other studies to extend to other times and places. That is Contextual in action.

(Volz et al 2020) has a narrower goal than (Ferguson et al 2020). It seeks to estimate the transmissibility of the then-new B.1.1.7 Covid 'variant of concern', i.e., what the WHO later designated the Alpha variant. All of its data are from England between October and early December 2020. Experiments were not feasible, so the paper conducts various observational studies, combining epidemiological and genetic data. It pursues five independent lines of analysis, each of which turns out to concur on roughly the same conclusion: that Alpha is more transmissible than the original variant, to the extent that it increases the virus's R-number (in England, in this period) by between 0.4 and 0.7. These five lines of analysis are:

1. The time and location of almost 2,000 Alpha cases from random population sampling was tracked, along with almost 50,000 non-Alpha samples. This revealed the increasing prevalence of Alpha relative to the original variant. To infer the growth difference per generation, a simple model was used of how the populations of two viruses with different R-numbers evolve. This model required the paper to estimate the virus's generation time (6.5 days) and the R-number at that time of the original variant (1.0).

2. The increasing prevalence of a particular genetic feature associated with Alpha – absence of the so-called S-gene – was traced, based on data from national positive Covid test results. This tracing was possible because, conveniently, almost a third of positive test samples in November and December 2020 recorded the presence or absence of the S-gene. The strength of association between Alpha and absence of the S-gene itself changed over time (because some non-Alpha variants also lack the S-gene), and this variation had to be modelled as a

[1] One of (Volz et al 2020)'s co-authors is Neil Ferguson, and several other of the co-authors, including Erik Volz, are also co-authors of (Ferguson et al 2020). (Some are co-authors of Brazeau et al 2020 as well.) So, my earlier criticisms of (Ferguson et al 2020) are certainly not ad hominem! The Imperial College unit is a great center of expertise. Nonetheless, as will become clear, I think the epistemological difference between (Volz et al 2020) and (Ferguson et al 2020) is stark, far beyond any difference acknowledged by their authors.

function of the date and area of the test sample. Overall, it turned out that the spread of Alpha inferred in 1 and 2 correspond closely.

3. The pattern of geographical expansion (as opposed to national prevalence) of absence of the S-gene was tracked, again using data from national positive Covid test results, and again inferring Alpha prevalence by means of the intermediary model mentioned in 2. The result was consistent with a greater transmissibility for Alpha, to a similar degree as calculated in 1 and 2.

4. A positive correlation was established between the estimated prevalence of Alpha, and independently derived estimates of the overall Covid R-number at different times and places. The paper ran a series of statistical regressions, using different measures of Alpha prevalence, different subdivisions of areas, and both frequentist and Bayesian estimation techniques. Quantitative estimates could be derived for the increase in the R-number associated with Alpha. These estimates were roughly consistent with those derived from the other lines of analysis.

5. A semi-mechanistic genetic model was fitted to the case numbers for Alpha and the original variant, to derive from many separate regressions further independent estimates of the R-numbers for each. These new estimates again roughly agreed with those derived from the other lines of analysis.

(Volz et al 2020) has two distinct goals. One is to estimate Alpha's greater transmissibility quantitatively. But a second, prior goal is to establish the qualitative claim that Alpha is more transmissible at all. Could Alpha's increased prevalence be explained in some other way?

The area-level data enabled one rival explanation to be ruled out, namely that Alpha might have a shorter incubation period. For a given transmissibility, a shorter incubation period leads a variant's case numbers to be more volatile: they increase more quickly than otherwise when $R > 1$, and decrease more quickly than otherwise when $R < 1$. But the data showed that, in areas where $R < 1$, Alpha cases did not decline faster than those of the old variant, contrary to the hypothesis of a shorter incubation period. On that hypothesis, random variation should also have meant Alpha cases going up and old variant cases going down in the same number of areas as the opposite pattern of Alpha cases going down and old variant cases going up. But in reality, there were plenty of areas with the former pattern but almost none with the latter – consistent with Alpha being more transmissible but not with it having a shorter incubation period. The rival explanation is disconfirmed.

(Volz et al 2020) also mentions a second rival explanation, namely that increases in frequency are due to chance rather than being more transmissible. This might be because of founder effects, which are especially likely if a variant is introduced from overseas. But unlike with some other variants whose prevalence increased, Alpha expanded from within England. In addition, the correlation between greater R-numbers and greater Alpha prevalence was observed in multiple regions. This tells against the chance hypothesis.

No other rival explanations are mentioned, and it is hard to think of plausible ones. The qualitative conclusion that Alpha is more transmissible is therefore well supported.

Overall, (Volz et al 2020) is excellent work. How does it succeed? First, it uses models that are relatively simple, and second, it uses them in a suitably contextual way. This allows it to be confident that its models are empirically accurate, as the details above show. In the paper's own words (18): 'We focused on relatively simple, data-driven analyses using parsimonious models making parsimonious assumptions, rather than, for instance, attempting to model the long-term transmission dynamics of [the Alpha and original variants] more mechanistically.' In other words, it did not follow the example of the CovidSim model – wisely. If more ambitious scope means loss of empirical confirmation, then it is a mistake.

To illustrate, consider the model used in (Volz et al 2020)'s first line of analysis. This model concerns how the relative prevalence of two viruses with different R-numbers changes over time, and it requires only two parameters to be estimated: the virus's incubation period, and the R-number of the original variant. The model is disanalogous to CovidSim in several other respects too. Most importantly, it has a history of empirical success. Examining such a history is the best way to learn in what conditions a model is likely to be successful again (Ylikoski 2019). The model captures the dynamics of virus reproduction when there are no complicating social factors. Experience gives us confidence that there was no significant interference by social factors in this case. Social factors do enter the model indirectly, via the value of the R-number, but there were good independent estimates of this R-number's value, and also good reason to think that that value was relatively constant across the relevant time period and geographical areas. The virus's incubation period was also well known independently, (Volz et al 2020) argues. And straightforward sensitivity analysis showed that the results were not unduly hostage to remaining small uncertainties.

For these reasons, confidence in this model is warranted here, unlike for CovidSim. Similar remarks apply to the other models used in (Volz et al 2020). And similar remarks apply also to many other excellent studies carried out during the pandemic. What matters each time is that a model is developed and used in accordance with Contextual rather than Master-Model: choice of model must be carefully justified each time, and empirical confirmation is crucial for that.

The important difference between the (Volz et al 2020) and (Ferguson et al 2020) papers is not that in the former models were used for retrospective estimation and in the latter for prediction. Rather, it is that the models in (Volz et al 2020) have empirical warrant whereas CovidSim does not, which in turn is because of the different methodological approaches.

There is an important caveat, though: in accordance with Contextual, predictions are warranted only when we have good reason to believe that the models that generated them still apply. (Volz et al 2020) acknowledges this (17): 'these estimates of transmission advantage apply to a period where high levels of social distancing were in place … extrapolation to other transmission contexts … requires caution.' This is wise, because whenever relations are unstable, extrapolation of results requires caution, and R-numbers are known to be sensitive to many environmental changes. We are not in Newton's world, so to speak. And indeed, it seems that Alpha's transmissibility advantage did not stay the same after the period of the study (Lemoine 2021a). Further, there is a repeated pattern in the pandemic of new variants enjoying large initial transmissibility advantages that subsequently diminish sharply, again suggesting caution when extrapolating initial calculations (Lemoine 2021b).

Some are skeptical that Alpha's transmissibility advantage is stable even within the narrow range of times and places that (Volz et al 2012) focuses on (Lemoine 2021b). One reason for that skepticism is a pattern of widely varying R-numbers across English regions at a given time, over and above measurement error. Against that, (Volz et al 2020)'s quantitative conclusions are obviously strengthened by the convergence of five independent lines of inquiry. This triangulation virtue is distinct from the paper's modeling virtues. Of course, those modeling virtues are what makes each line of inquiry persuasive individually, and if component lines of inquiry are not persuasive individually then it is dubious that triangulation retains its epistemic force (Odenbaugh and Alexandrova 2011, Betz 2015). And perhaps there is convergence merely on the contemporary average R-number in England, hiding local variation. For our purposes, what matters is that, regardless of the exact credence given to (Volz et al 2020)'s quantitative conclusion, it is clear that that credence should be proportional to the extent to which the conclusion enjoys local empirical support, which in turn depends on the extent to which it is based on methods that follow Contextual.

4.2. Informal methods

Sometimes, the only methods available are informal ones. Consider border controls again. Their impact on infection rates varies, and how infection rates are impacted by other things varies too. The same is true of other target variables besides infection rates. These instabilities mean that master models are ill-suited to evaluating the impact of border controls. (I am not aware of any actual such model; the CovidSim model, recall, omitted border controls altogether.)

In which case, what to do? Only one option remains, namely, no formal model at all. This amounts to causal reasoning based on evidence, including quantitative evidence, but done in the manner of careful historians or qualitative social scientists. The 'models' in these cases may be no more than simple causal claims, such as "border controls reduce Covid cases". More realistically, the causal claims will be more detailed: "if border controls are organized in manner X at stage Y of the pandemic, then in countries with a high throughput of travelers they reduce Covid cases by Z". How to confirm such claims? In effect, we approximate natural experiments as best we can, or make single-case causal inferences that rely on background knowledge to evaluate the implicit counterfactuals. (Grépin et al 2020, for example, reviews early work on border controls to support the claim that, at the beginning of the Covid pandemic, travel restrictions around Wuhan reduced the importation of cases internationally – contrary to the experience of influenza outbreaks.) Just because these methods are more informal does not mean they are not empirical. To be sure, inference might be more difficult when we cannot use methods such as controlled experiments. But the perfect should not be the enemy of the good. Informal methods can still carry warrant, on pain of being a complete skeptic about large areas of scientific – and indeed everyday – inquiry. It is by informal methods that best practice about border controls has been established and shared, such as operational details of hotel quarantine, or how border requirements should vary depending on a traveler's country of origin, or how their impact varies depending on current case numbers and on whether those numbers are rising or falling. Informal models, and informal confirmation procedures for them, are certainly preferable to the CovidSim model. The key epistemic criterion is the same for all models alike, formal or informal: empirical confirmation.

Similar remarks apply widely. Consider the vexed issue of lockdowns: what is their true impact on infection numbers, economic output, and mental health? There has been a huge amount of work about this, of course, and reviewing it is beyond the scope of this paper. I mention lockdowns only to point out several ways in which this paper's discussion bears on the matter. First, likely there is no univocal answer. The impact of lockdowns varies with the exact lockdown regulations being imposed, by whom, on what community, and at what stage of the pandemic. During a recession, during winter or summer, in an urbanized or rural country, in a rich or poor one? In a country with lots of gig workers, with lots of multi-generational households, with a history of suspicion of government? On a population that is fit, with low levels of co-morbidities, with access to high ICU capacity and extensive primary care? Even within a single country – the UK – the impact of the same lockdown regulations, at the same stage of the pandemic, showed big differences across regions and sectors. Second – which follows from the first – Contextual methods are required. The impact of lockdowns should not be assessed by applying a single master model such as CovidSim (even though this was what actually happened with the original lockdown decision in the UK). Third, might there nonetheless be some rule of thumb that is relatively stable, perhaps that lockdowns' effects have been overrated or underrated? Perhaps, perhaps not. The Contextual methodology says that that question can be answered only by close empirical analyses, in the manner of (Volz et al 2020).

Because different countries vary so much in relevant ways, it would not be helpful to run a statistical regression across them to assess the impact of lockdown policy. Cross-country comparisons need to be more nuanced. For the same reason, a hypothetical randomized trial would tell us little. Methods such as regressions and trials are very effective – but only when the target relations are stable. That's why they work well for assessing vaccines but not for assessing lockdowns.

Similar remarks apply to many other policy questions: how much to test, what mask-wearing and social distancing to require, whether to close schools and universities, how to regulate those asked to self-isolate, how to contact-trace, and how to allocate medical equipment. What can be learned by comparing the experiences of different countries? Which local details matter, and which don't? Informal methods are the way to find out (Han et al 2020).

How might informal methods have tackled (Ferguson et al 2020)'s original task, namely to project the number of infections under various policy scenarios? Such methods could have been applied to assess Covid developments in China, Italy, and other countries, experience of previous epidemics, and background knowledge of local health systems and political cultures. Even at the beginning of the pandemic, there was plenty of such evidence (Lipsitch 2020). Still, most likely only very rough projections could be justified initially. If so, it is better to accept that truth than to imagine that more precise projections are trustworthy. If a model is unwarranted then it is not useful, and that does not change just because an epistemic situation is difficult and there are few alternatives.

There is continuity between informal methods and the methods of (Volz et al 2020). How formal a model is, and how formal the techniques are by which to test whether a model applies, may vary. But both informal methods and (Volz et al 2020) share the same Contextual methodology, and that is what matters.

Recent work in philosophy of science emphasizes 'middle-range theories' (Cartwright 2020): such theories are, roughly speaking, models of non-universal scope that are applied contextually in conjunction with local, often informal knowledge. This too is an instance of the Contextual methodology.

5. Conclusion

The Master-Model strategy does not work for the Covid pandemic. No candidate master model boasts the needed empirical success, and likely none will: the real-world targets are too unstable. A different modeling strategy is required. The same conclusion will apply whenever target relations are fragile, which likely means pandemics generally.

This conclusion is not of mere ivory-tower interest. Pursuing Master-Model is positively harmful when it diverts resources away from a superior Contextual alternative. Facing a policy emergency in March 2020, the initial contrast was between relying on models such as CovidSim versus relying only on informal methods. At the beginning of the pandemic, SAGE was criticized for being top-heavy with mathematical modelers rather than empirical field scientists (Ford 2020, Costello 2020). By the time this began to be (slightly) rectified, fateful policy mistakes had already been made: relative neglect of on-the-ground experience from other countries and from practitioners at home, is widely alleged to have slowed the provision of protective equipment to health workers, and to have slowed the setting up of a testing system.

We know a lot more now than at the beginning of the Covid pandemic, so both projections and policy responses are far better grounded than they were. But this welcome progress has not come from grand, one-size-fits-all models. Rather, it has come from a huge accumulation of knowledge gained by informal methods and by modeling that is empirically confirmed. That is the way forward. Contextual was, and is, required.

References

Alexandrova, A. (2008). ‘Making models count’. *Philosophy of Science* 75, 383-404.

Alexandrova, A., and R. Northcott (2009). ‘Progress in economics: Lessons from the spectrum auctions’, in H. Kincaid and D. Ross (eds.), *The Oxford Handbook of Philosophy of Economics*, Oxford University Press, 306-337.

Aydinonat, E. (2008). *The Invisible Hand in Economics: How Economists Explain Unintended Social Consequences* (London: Routledge).

Betz, G. (2015). ‘Are climate models credible worlds? Prospects and limitations of possibilistic climate prediction’, *European Journal for Philosophy of Science* 5.2, 191-215.

Billah, A., M. Miah, and N.Khan (2020). ‘Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence’, *PLoS ONE* 15.11: e0242128. https://doi. org/10.1371/journal.pone.0242128, 11th November 2020.

Brazeau, N., R. Verity, S. Jenks, H. Fu, C. Whittaker, P. Winskill, I. Dorigatti, P. Walker, S. Riley, R. Schnekenberg, H. Hoeltgebaum, T. Mellan, S. Mishra, J. Unwin, O. Watson, Z. Cucunuba, M. Baguelin, L. Whittles, S. Bhatt, A. Ghani, N. Ferguson, and L. Okell (2020). ‘Report 34 – COVID-19 Infection Fatality Ratio Estimates from Seroprevalence’, *MRC Centre for Global Infectious Disease Analysis*, 29th October 2020.

Broadbent, A. (2013). *Philosophy of Epidemiology* (Palgrave MacMillan).

Cartwright, N. (1989). *Nature’s Capacities and Their Measurement*. (Oxford: Oxford University Press.)

Cartwright, N. (2019). *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. (Open Court Publishing.)

Cartwright, N. (2020). ‘Middle-range theory: Without it what could anyone do?’, *Theoria* 35.3, 269-323.

Conn, D., F. Lawrence, P. Lewis, S. Carrell, D. Pegg, H. Davies, and R. Evans (2020). ‘Revealed: the inside story of the UK’s Covid-19 crisis’, *The Guardian*, 29th April 2020. https://www.theguardian.com/world/2020/apr/29/revealed-the-inside-story-of-uk-covid-19-coronavirus-crisis

Costello, A. (2020). ‘The government’s secret science group has a shocking lack of expertise’, *The Guardian*, 27th April 2020. The government's secret science group has a shocking lack of expertise | Coronavirus | The Guardian

Edeling, W., A. Hamid, R. Sinclair, D. Suleimenova, K. Gopalakrishnan, B. Bosak, D. Groen, I. Mahmood, D. Crommelin, and P. Coveney (2020). ‘Model uncertainty and decision making: Predicting the Impact of COVID-19 Using the CovidSim Epidemiological Code’, 10.21203/rs.3.rs-82122/v1.

Ferguson, N., D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. Okell, S. van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. Walker, C. Walters, P. Winskill, C. Whittaker, C. Donnelly, S. Riley, A. Ghani (2020). ‘Report 9 – Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand’, *MRC Centre for Global Infectious Disease Analysis*, 16th March 2020.

Forber, P. (2010). ‘Confirmation and Explaining How Possible’, *Studies in History and Philosophy of Science Part C* 41, 32-40.

Ford, J. (2020). ‘The battle at the heart of British science over coronavirus’, *Financial Times*, April 15th 2020. https://www.ft.com/content/1e390ac6-7e2c-11ea-8fdb-7ec06edeef84?desktop=true&segmentId=7c8f09b9-9b61-4fbb-9430-9208a9e233c8

Fuller, J. (2021). ‘What are COVID-19 models modeling (philosophically speaking)?’ *History and Philosophy of the Life Sciences* 43, 47 https://doi.org/10.1007/s40656-021-00407-5

Grépin, K., T. L. Ho, Z. Liu, S. Marion, J. Piper, C. Worsnop, and K. Lee (2020). ‘Evidence of the effectiveness of travel-related measures during the early phase of the COVID-19 pandemic: a rapid systematic review’, *BMJ Global Health* 6:e004537 doi:10.1136/bmjgh-2020-004537

Grey, S., and A. MacAskill (2020). ‘Special Report: Johnson listened to his scientists about coronavirus - but they were slow to sound the alarm’, *Reuters*, April 7th 2020. Special Report: Johnson listened to his scientists about coronavirus - but they were slow to sound the alarm | Reuters

Grüne-Yanoff, T. (2009). ‘Learning from Minimal Economic Models’, *Erkenntnis* 70, 81-99.

Han, E., M. Tan, E. Turk, D. Sridhar, G. Leung, K. Shibuya, N. Asgari, J. Oh, A. García-Basteiro, J. Hanefeld, A. Cook, L. Hsu, Y. Teo, D. Heymann, H. Clark, M. McKee, and H. Legido-Quigley (2020). ‘Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe’, *Lancet* 396, 1525-34.

Lemoine, P. (2021a). ‘The British variant of SARS-CoV-2 and the poverty of epidemiology’, CSPI blog, 9th April 2021 https://cspicenter.org/blog/waronscience/the-british-variant-of-sars-cov-2-and-the-povery-of-epidemiology/

Lemoine, P. (2021b). ‘Is the Delta variant really more than twice as transmissible as the original strain of the virus?’, CSPI blog, 31st August 2021 https://cspicenter.org/blog/waronscience/is-the-delta-variant-really-more-than-twice-as-transmissible-as-the-original-strain-of-the-virus/

Lipsitch, M. (2020). ‘We know enough now to act decisively against Covid-19. Social distancing is a good place to start’, *Stat News*, March 18th 2020. We know enough now to act decisively against Covid-19 - STAT (statnews.com)

Mäki, U. (1992). ‘On the method of isolation in economics’, Poznan Studies in the Philosophy of the Sciences and the Humanities 26, 19-54.

Manzo, G. (2020). ‘Complex Social Networks are Missing in the Dominant COVID-19 Epidemic Models’, *Sociologica* 14.1, 31-49.

Mill, J. S. (1843). *A System of Logic*. London: Parker.

Northcott, R. (2013). ‘Degree of explanation’, *Synthese* 190.15, 3087-3105.

Northcott, R. (2017). ‘When are purely predictive models best?’ *Disputatio* 9.47, 631-656.

Northcott, R. (forthcoming). *Science for a Fragile World* (Oxford: Oxford University Press.)

Northcott, R., and A. Alexandrova (2013). ‘It’s just a feeling: why economic models do not explain’, *Journal of Economic Methodology* 20, 262-267.

Odenbaugh, J., and A. Alexandrova (2011). ‘Buyer Beware: robustness analyses in economics and biology’, *Biology and Philosophy* 26, 757-771.

Schroeder, A. (2021). ‘How to interpret COVID-19 predictions: reassessing the IHME’s model’, *Philosophy of Medicine*, 2(1), 1–7.

Teller, P. (2001). ‘Twilight of the perfect model model’, *Erkenntnis* 55.3, 393–415.

van Basshuysen, P., and L. White (2021a). ‘Were lockdowns justified? A return to the facts and evidence’, *Kennedy Institute of Ethics Journal* 31.4, 405-428.

van Basshuysen, P., and L. White (2021b), ‘The epistemic duties of philosophers: an addendum’, *Kennedy Institute of Ethics Journal* 31.4, 447-451.

Veit, W. (2019). ‘Model pluralism’, *Philosophy of the Social Sciences* 50.2, 91–114.

Volz, E., S. Mishra, M. Chand, J. Barrett, R. Johnson, L. Geidelberg, W. Hinsley, D. Laydon, G. Dabrera, Á. O’Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. Ariani, O. Boyd, N. Loman, J. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. Kwiatkowski, The COVID-19 Genomics UK (COG-UK) consortium, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, and N. Ferguson (2020). ‘Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England’, *Nature* 593, 266-269.

Winsberg, E., J. Brennan, and C. Surprenant (2020). ‘How Government Leaders Violated Their Epistemic Duties During the SARS-CoV-2 Crisis’, *Kennedy Institute of Ethics Journal* 30.3-4, 215-242.

Winsberg, E., J. Brennan, and C. Surprenant (2021). ‘This Paper Attacks a Strawman but the Strawman Wins: A reply to van Basshuysen and White’, *Kennedy Institute of Ethics Journal* 31.4, 429-446.

Ylikoski, P. (2019). ‘Mechanism-based theorizing and generalization from case studies’, *Studies in the History and Philosophy of Science* 78, 14-22.