Goodhart's law and the dark side of herd immunity

Herd immunity is often held up as a solution to the COVID-19 crisis. ***Bob Hancké (LSE)*** *argues that it is a dangerous solution, and morally rejectable – in large part because it is a special instance of Goodhart's law, undermining the very goal it purports to achieve. Herd immunity is not only technically flawed, as many medical experts argue, but also epistemologically wrong in the case of COVID-19.*

In mid-October a group of respected medical scientists issued the (somewhat pompously titled) Great Barrington Declaration (GBD), which urged, among other things, the adoption of a 'herd immunity' strategy to cope with COVID-19. The economic, social and physical and mental health costs of severe constraints on normal activity are now larger than they would be in a situation where the pandemic could just run its course. The GBD recommends that a sizeable share of the population build up immunity while protecting the most vulnerable, thus stopping the growth of the viral infection.

While I am not an epidemiologist or in any other way trained in matters related to health, I do teach research design and care deeply about the quality of public policy, especially in the case of a nasty epidemic like COVID-19. As the debate about herd immunity, especially after the GBD, gathered steam, a problem that has haunted many well-meant policies suddenly seemed to make a guest appearance. Made famous by LSE professor Charles Goodhart, the law with his name goes roughly like this: *If an observed (statistical) regularity becomes the target of an intentional policy, it ceases to be a meaningful measure for policy making.*

In what follows, I start by unpacking the concept of herd immunity in light of the ironies associated with Goodhart's law and then develop a very simple model to assess it as a strategy for coping with COVID-19. The upshot is this: the law is alive and well in this area; if adopted, herd immunity would not only undermine its own targets, but its death toll is way beyond what modern societies should or could countenance.

The immunity of the herd

Photo: Qtii via a CC-BY-NC-SA 2.0 licence

Date originally posted: 2020-10-23

Permalink: https://blogs.lse.ac.uk/covid19/2020/10/23/goodharts-law-and-the-dark-side-of-herd-immunity/

Blog homepage: https://blogs.lse.ac.uk/covid19/

Herd immunity is – for those very, very few among us lucky enough not to have been bombarded with it since early spring 2020 – an epidemiological term referring to the level of aggregate immunity in a population that is necessary to stop an epidemic. The basic idea is that if a certain share of the population, usually held to be around 75%, is immune to a virus, its chances of spreading in the population fall to almost zero. It tries to go from patient A to B, but B has a 75% chance of being immune, so it is very likely not to make it past B; aggregate this over the entire population and the net effect is that almost all contagious jumps from A to anyone else will fail and the virus will slowly die out. While there are loads of possible qualifications (speed of transmission, the immunity rate fluctuating between high and low intensity periods, etc.), none of these would change the fundamental logic. Importantly, you do not need 100% coverage for herd immunity to kick in. Everything above 60% is a good start (in 3 cases out of 5 the disease will not be transmitted), and 75% (3 out of 4) is almost certain to do the trick entirely.

Like all ideas, the notion of herd immunity is based on a few assumptions – most of which are related to its origin in veterinary medicine, where it is a technique to counter a contagious infectious disease making its way through the relatively small local bovine population. Vaccinating a minimum threshold number will create the positive externality explained earlier, in which even non-vaccinated cows run a very low risk of infection. It is also used in mass vaccination programmes for humans, but usually in a quite tightly controlled environment where the same assumptions have to hold: you need to know everything relevant on the evolution of the contagious disease, be certain that immunity will result from the treatment, and be able to measure the number of vaccinations, the degree of immunity and the size of the risk population accurately. Sadly, thus far COVID-19 has escaped much of our understanding in all or most of these areas.

'We know nothing'

Not only do we know relatively little about the evolution of the disease, as an excellent article in The Atlantic explains, we have no idea, really, what exactly drives the spread of the coronavirus and what the best mitigation strategies are. For example, why did a few small towns in northern Italy record more deaths than the rest of the country combined? Or why did South Korea manage to control the spread of coronavirus, with only a few hundred excess deaths after a very inauspicious start? Apparently, the R number that everybody has focused on to understand the coronavirus's road through the population is one measure, but probably not a very helpful one. Dispersion rates – and overdispersion as a super-spreader handmaiden – are at least as important, but little is known about how to handle let alone prevent them. In short, we do not understand COVID-19 all that well. Effective vaccines that can be manufactured at scale are, if Ebola, HIV and a few other viruses are anything to go by, possibly still quite a way off. Most of the conditions for a successful herd immunity strategy, and especially the basic knowledge, are simply absent, as an open letter in The Lancet argues.

The dark side of herd immunity

But there is more. The notion of herd immunity produces a series of paradoxes that actually undermines the very idea itself. As a strategy it lacks precision; as a technique it is brutal; and dynamically it is self-destroying. Take each one of these points at a time.

Starting with the basics, let us assume, generously, that once recovered a person is fully immune. It is not certain if this is the case with COVID-19 – there are reported cases of reinfection within one cycle – but assuming immunity after surviving the infection simply stacks the cards against our argument (and if this were not true, the idea of herd immunity makes little sense), so no harm done from a logical and methodological point of view.

Then define herd immunity (H) as the ratio of living immune (R) to P, the total living population: H=R/P. For the sake of the argument, we assume that the target value for H is 75%, i.e. three people out of four need to have been infected and then become immune (or were, for a variety of unknown other reasons, naturally immune). Again, given what we know about COVID-19 in particular and highly contagious viral diseases in general this may not be a high enough value for H; but the H=0.75 assumption simply means that we stack the cards *against* our argument a second time, without producing a problem for our model. Methodologically these two assumptions are akin to what is known as a critical or limiting case strategy: if our argument holds under these adverse conditions for it to work, the point we make will certainly be true under more favourable circumstances.

Date originally posted: 2020-10-23

Permalink: https://blogs.lse.ac.uk/covid19/2020/10/23/goodharts-law-and-the-dark-side-of-herd-immunity/

Blog homepage: https://blogs.lse.ac.uk/covid19/

An outcome, not a strategy

Back to the H value, now. The interesting thing about this ratio (and all others) is that it is a *compound variable*, with two *direct variables* of interest at its roots – in this case R and P – but with an indeterminate relationship. To reach the target H value, R can increase while P remains stable; P can decrease with R stable; both can move simultaneously in opposite directions, R up and P down; both can increase but R increases more than P; or both can decrease but P decreases faster than R. All these combinations will lead to a rise in H. Nothing about H tells us *a priori* what needs to happen: herd immunity is an outcome of two analytically and practically separate processes, both of which can, at best, only be imperfectly controlled and monitored. This raises the first problem with the concept of herd immunity: since it is a compound variable, the outcome of two complex processes, it is not obvious how it can be a clear, deliberate strategy.

Secondly, the idea has some perverse consequences as a result of this lack of conceptual clarity, as the simple model below illustrates. If A infects B and B dies, the absolute number in R remains constant, while the absolute number for P goes down by 1. That has two very different effects for the ratio R/P. The first is that as a result of B's death, one infection did not result in a person who is immune, a gap that can only be filled by another infected-and-then-immune person (call this *substitution*). The second effect, however, pulls in the opposite direction: with a smaller population (Pt0-1 death), we need, in absolute terms, fewer people to be infected to get to the 75% target value for H (call this subtraction).

More, not fewer, infections

Here a subtle version of Goodhart's law comes into play. Because the effect of the target associated with substitution (0.75) is smaller than the target associated with subtraction (1), deaths as a result of infections *increase* the number of people who need to be infected to get to 75% of the population. In plain English: trying to bring the infection under control means that more people have to be infected.

As if that was not enough, there's a morbid twist to this dynamic: because infections carry with them the risk of death, herd immunity leads to a non-trivial increase in the number of additional deaths above those to be expected.

To illustrate this, let us use some actual numbers (I borrowed these from the sister post by Nick Barr): assume H to be 75% and the death rate for the infection to be 1%. In a country like the UK (population c. 65 million) reaching an H value of 75% requires that a whopping 48.75m need to be infected (note that the UK is currently estimated to have a 6% immunity rate, and even if it were double that, reaching a H value of 75% would remain a steep hill to climb). Of those 48.75m, 1% or 487,500 will die. (Interestingly, this is almost exactly the estimated number of deaths predicted by the Imperial College model as the basis of the UK's spring 2020 lockdown, if the virus had not been countered with stringent containment measures.) If we get the death rate slightly wrong and it is 2%, 975,000 people have to die to reach a H value of 75%. Gulp.

The grim reaper's second bite of the cherry

The story gets better (or, actually, worse). To achieve an overall infection survival rate of 75%, it is necessary to replace some of the initially infected who died, i.e. to expose more people to infection, as we said earlier. But that is not without costs. With a 1% death rate, a target H value of 75% requires an additional 0.18% exposure, in fact. Assuming a 1% death rate, the number of deaths rises to 488,722 (and over 977,000 in the case of a 2% death rate), i.e. an additional 1,222 (or 2,444) deaths over and above the already staggering baseline numbers.

In sum, there are many very serious problems with the concept of herd immunity. First, conceptually it is hard to think of herd immunity as a clear strategy, given that it consists of two imperfectly understood variables that are only partially under our control. Secondly, a sustained herd immunity strategy kills, *prima facie*, at least 487,000 people (I abstract here from improvements in health care, which might flatten the curve somewhat but won't fundamentally change the dynamic unless there is widespread access to an effective vaccine). Finally, the perverse effect associated with this application of Goodhart's law adds an additional 1,200 (or up to 2,500 if the death rate is above 1%). Such numbers are the stuff of mid-20th century horrors, dystopian movies and end-of-time sci-fi novels, not of sensible policy making.

Date originally posted: 2020-10-23

Permalink: https://blogs.lse.ac.uk/covid19/2020/10/23/goodharts-law-and-the-dark-side-of-herd-immunity/

Blog homepage: https://blogs.lse.ac.uk/covid19/

'Protecting' the vulnerable

There is only one logically tight counterargument: protect, i.e. isolate, the vulnerable part of the population with known co-morbidities. That makes some logical sense, in the same way that you could theoretically imagine a single wee-free swimming pool lane. The problem with the idea is that it stumbles at the first practical hurdle: even assuming the vulnerable population is correctly identified (a big if, considering how little we know), how do you stop them from starving, shopping, talking to neighbours, family, etc. for the possibly very long time it takes to get to a 75% H value? At a rate of increase of 7% per annum, roughly the figure for 2020, we are talking about a decade of 'protection'. And even with an effective vaccine many will still face several years of such 'protection'. The New York Times ran an eyebrow-raising article on that particular problem with 'mass murder' in the title.

Bull immunity and his excrement

Carl Bergstrom and Jevin West, the authors of the highly insightful and entertaining Calling Bullshit: The Art of Skepticism in a Data-driven World refer to Goodhart's law as a heuristic to spot BS – not unlike what I have done here. Their last chapter, 'Refuting Bullshit', invites all of us to point the finger at BS when we spot it. In that spirit, I hereby declare herd immunity in today's situation a dangerous technocratic fool's errand, without any basis in fact or science. Bullshit, in other words.

This post represents the views of the author and not those of the COVID-19 blog, nor LSE. The author wishes to thank Nick Barr, Henrike Granzow and Laurenz Mathei for helpful comments on an earlier draft.