A simple introduction to epidemiological modelling—the SIR model

Alexander Bird

King's College London
Peter Sowerby Philosophy & Medicine Project

1 Introduction

Epidemiological models have been frequently mentioned in the media lately. What are they? And how do they work? In this I will focus today on the model that is the simplest and most frequently used model in epidemiology, the SIR model. This model was developed by Kermack and McKendrick in 1927 and has its origins in the work of Sir Ronald Ross, who won a Nobel prize for his research on the transmission of malaria.

2 What is a scientific model?

First, what is a scientific model? A model is a picture of what something is or of how it works that is simplified when compared to reality. In a good model that simplification is an advantage because it can reveal the basic structure or mechanism of what is going on and thereby give us a deeper understanding of the thing we are interested in. If the model is a mathematical, then the simplification may also allow us to use mathematics to understand and predict things in a way that would not be possible with a more complicated model. Often the simplified model is all we need for our purposes. But a good model also allows itself to be developed and made more sophisticated in order to describe and understand further details as necessary. This is the case with the SIR model.

Think of an architect's scale model of a building. This will omit a lot of detail, such as the decorative ornaments. But it will serve the important functions of giving a good idea of what the building will look like when built and how it will fit with neighbouring buildings. It might reveal the basic engineering issues that need to be faced in construction, and so forth. You may remember being taught the model that treats a gas as being made up of millions of spherical particles of negligible size that bounce of each other and the walls of the container they are in. This 'ideal gas' model is simplified, since gas molecules are not spherical, and they do not just bounce off each other, they have faint forces of attraction between them, and in a high pressure gas their size is not negligible, and so on. But for understanding the basic reasons why pressure and temperature in a gas increase when you compress it, this model is enough. And its predictions are good enough for many purposes.

An important point to make about models is that they typically must make certain assumptions. Sometimes these are assumptions about the world that the mod-

ellers think are true. It will be the job of scientists to find out whether they really are true (or are true in the case to which the model is being applied). Sometimes, the modellers know that their assumptions are not strictly true—they are made as part of the simplifying process. But, in a good model, these assumptions will not be far from the truth, so the model still gives us a good understanding of what is going on and may give us reliable predictions.

3 The SIR model

The epidemiological model we will look at is the SIR model. 'SIR' stands for Susceptible–Infectious–Recovered'. It aims to give us an understanding of the relationship between the number of people infected with a disease and the numbers who are uninfected and still susceptible and those who have recovered and are no longer susceptible. So the total population is divided into these three groups, or 'compartments' as they are called and we want to know how fast people move between them, from susceptible to infectious, and from infectious to recovered over the course of an epidemic. This may allow us to predict the course of an epidemic or to intervene to slow it down or even to prevent it.

We can see already that our model makes certain assumptions. It assumes that the overall population stays constant—there are not people being born and, more relevantly perhaps, dying. For current purposes that does not matter, since over the course of a few months, the numbers of births and deaths is not very large compared to the overall population. Another assumption is that the infected people and the infectious people are the same. Again, that is not always true, because there can be a latent period after becoming infected but before being infectious to others. For current purposes it is okay to assume that there isn't a significant latent period.

Let us call susceptible people, 'susceptibles' and the number of them we call 'S'. Let us call infected and infectious people, the 'infectives', and the number of them is 'I'. And the number of recovered people, the 'recovereds', is 'R'. The total population is N, so $N = S + I + R$ (i.e. everyone is either susceptible, or infected, or recovered).

We start by asking how many of those who are susceptible will become infected in one day. Think of a single infected person who comes into contact with n other people in a day in a way that could potentially infect them. And now think of the chance that one such encounter will lead to a new infection. We use the Greek letter 'τ' for the probability that an encounter between an infectious person and a susceptible leads to the latter becoming infected. So if all the n people are susceptible, the number of people infected in a day by our single infective is τn. This is an important number—the number of people an average infected person will infect if everybody else is susceptible—and usually we use the Greek symbol β to denote it. For example if 'τ' is 0.25, meaning that any encounter has a probability of $1/4$ of leading to a new infection, and n is 8, meaning that the infectious individual encounters 8 other, susceptible people, then there will be 2 new infections that day, so $\beta = 2$. It is intuitively obvious why diseases spread quickly in cities such as London and New York, where many people use crowded public transport to commute to work. A Londoner or New Yorker will come into physical proximity to a large number of people, meaning a high n. And τ may be high because those encounters may have a higher chance of passing on the disease—think of people rubbing shoulder-to-shoulder on a tube or subway train.

Let's think a little more about τ, n, and β. τ is partly a biological matter—how contagious the disease it. But it will also depend on social factors. In some societies people engage in a greater degree of physical contact when they meet. In some circumstances there will be better opportunities for personal hygiene. These factors will make a difference to the value for τ in those societies. n is the average number of encounters a person has of a kind that would allow for the transmission of the disease. Regarding measures to combat the spread of COVID-19, hand-washing and wearing masks work to reduce τ while the lockdown and isolation reduce n. That said, the idea that τ and n are entirely distinct is somewhat doubtful. Consider the social distancing practice of not coming within 2m of other people. On the face of it this reduces n. If 2m were some kind of hard boundary, so that beyond 2m the disease cannot be passed on but within 2m it might be, then keeping people beyond 2m is a way of reducing the number of encounters that might spread the disease. On the other hand, more realistically, the 2m rule doesn't operate quite like that, since there can be transmission beyond 2m. This social distancing measure can then be thought of as reducing the chance of transmission. In which case this measure looks more like a reduction in τ. For this reason, while τ and n are useful in setting up the model, it is better on the whole to focus on β, the number of new infections an infectious individual would bring about each day when everyone else is susceptible.

We want to know the number of new infections in the population as a whole. We said that each infective infects β susceptibles each day. And there are I infectives in the population, each bringing about those β infections. So the number of new infections is βI.

That's on the assumption that everybody an infective meets is susceptible. Which will be true or close to the truth early on in an epidemic. However, in the middle of an epidemic not all the people in infective meets will be susceptible. Some will be immune because they have recovered. Let us return to our single infective for a moment. They encounter n people each day. Some of these will be susceptible, but not all. What proportion will be susceptible? We take that proportion to be the same as the proportion of susceptible people in the population as a whole, which is S/N. So if 50% of the population are already immune, then in out example although our infective encounters eight people, only four of then are susceptible and so given τ is $1/4$, there will be only one person newly infected each day.

That means in what we have already said we must replace n by nS/N, and so β by $\beta S/N$ So:

$$\text{Daily number of new infections} = \beta\frac{IS}{N}$$

The new infections all come from the susceptible portion of the population. So we can say:

$$\text{(I) Daily change in number of susceptibles} = -\beta\frac{IS}{N}$$

The minus sign is there to show that the change in susceptibles is a decrease. Each day, $\beta IS/N$ people become infected and so are no longer among the susceptibles.

Each day some of those who are infected recover. We use another Greek letter 'ν' (*nu*) to denote the proportion. (Sometimes γ (gamma) is used rather than ν. I regard ν as more appropriate as ν is the letter standardly used in physics and other disciplines to denote a frequency, which is what we are dealing with here. ν is the frequency—or probability—with which a single infective recovers in a day. This will become clearer later when we see that the reciprocal of ν is the average duration of

the disease in an infected individual.) So:

$$\text{(II) Daily change in number of recovereds} = \nu I$$

ν is the proportion of infected people who recover each day. One way to understand ν is to see that it is related to the duration of the illness. If it takes four days from being infected to recovering, then we would find that each day one quarter of those infected would recover. If it take ten days to recover, then each day only one tenth will recover. So ν is inversely related to the duration of the disease, D:

$$\nu = 1/D$$

The duration, D, is going to depend primarily on the biological nature of the illness. It will also depend on the medical care available: are there drugs that will help cure the illness? On the other hand, some treatments that help with symptoms in fact increase the duration of an illness. Fever-reducing medicines can extend the time to recovery because fever is one of the body's means of combatting infection. We can work out the value of D and so of ν by looking at the medical records of infected people.

How does the number of infectives change? That's the number of susceptibles who have become infected minus the number of infectives who have recovered:

$$\text{(III) Daily change in number of infectives} = \beta\frac{IS}{N} - \nu I$$

Mathematicians express changes using derivatives. The rate of the change of the number of susceptibles over time is expressed by:

$$\frac{dS}{dt}$$

So we can write our equations (I)–(III) thus:

$$\text{(I)} \quad \frac{dS}{dt} = -\beta\frac{IS}{N}$$

$$\text{(II)} \quad \frac{dR}{dt} = \nu I$$

$$\text{(III)} \quad \frac{dI}{dt} = \beta\frac{IS}{N} - \nu I$$

Which is the set of differential equations at the heart of the SIR model. It is not mathematically straightforward to solve these equations, since they are non-linear. But we can use them nonetheless to show exactly how S, I, and R change over time, depending on the values of β and ν.

4 The basic reproduction number $\mathscr{R}_0$

Let's look at the equation telling us how the number of infectives changes:

$$\text{(III) Daily change in number of infectives} = \beta\frac{IS}{N} - \nu I$$

If this number if positive, then the number of infected people is increasing, and the epidemic is growing—more people are becoming infected than are recovering.

The epidemic will be on the wane when this number is negative—when there are more people recovering than there are becoming infected, i.e. when:

$$\beta \frac{IS}{N} < \nu I$$

which we can simplify by dividing by I by rearranging S, N, and ν

$$\frac{\beta}{\nu} < \frac{N}{S}$$

The number β/ν is very important in epidemiology. As we can see, it is key to telling whether an epidemic is waning—or growing. It tells us how rapidly a new epidemic will grow. It can also describe the success of efforts to contain an epidemic. And it can be used to tell us how many people need to be immune in order for there to be herd immunity, as we will see shortly. Epidemiologists call it the *basic reproduction number* or the *basic reproductive number*. It is given the symbol $\mathscr{R}_0$. (I am using a curly '$\mathscr{R}$' to distinguish it clearly from the 'R' for recovereds, but usually a normal 'R' is used: R_0.)

So this equation:

$$\mathscr{R}_0 = \frac{\beta}{\nu}$$

defines $\mathscr{R}_0$. (Shortly we will look at an alternative but equivalent definition of $\mathscr{R}_0$.)

We can say therefore that the epidemic is waning when:

$$\mathscr{R}_0 < \frac{N}{S}$$

This I call the 'waning rule' for epidemics.

5 Understanding $\mathscr{R}_0$ and the waning rule

A more intuitive route to grasping what $\mathscr{R}_0$ means requires remembering that ν is equal to $1/D$, where D is the duration of the infectious period. So we can also define $\mathscr{R}_0$ thus:

$$\mathscr{R}_0 = \beta D$$

Remember that β is the number of new infections brought about each day by a single infective (assuming everyone else is susceptible). D is how many days they are infectious for. So βD is the total number of new infections the infective brings about before recovering and losing infectiousness. For example, let β be 2/3, so each day our single infective infects, on average to-thirds of a person. And imagine they are infectious for three days. By the end of their infectious period they will have infected two people. So this is a good way to understand $\mathscr{R}_0$—it is the number of new infections brought about, on average, by a single infective before they recover.

At this point it is important to distinguish discussion of an epidemic early on, when the number of infectives and recovereds is small compared to the total population, almost all of whom are susceptible, from discussion of an epidemic later on, when a substantial number of people have been infected and have recovered or, possibly, died.

Very often we are interested in the features of an epidemic in its early phase, when we wish to know what to expect and how to combat the spread of the disease.

In the early days or weeks of an epidemic even if the number of infectives and recovereds is in the thousands, it remains the case that in a country where the total population numbers in the tens of millions, almost everyone is susceptible. So S can be regarded as very close to N, and so $N/S \approx 1$. Therefore, early in the epidemic, we can say that total number of infections is decreasing and the epidemic is waning, when:

$$\mathscr{R}_0 < 1$$

If, on the other hand, $\mathscr{R}_0$ is greater than 1, the epidemic is continuing to spread.

Therefore we can also write the waning rule as:

$$\beta D < 1$$

Consider the individual infected person who infects β people in a day. He is infectious for D days. Therefore, overall he infects, on average, βD people. At the end of D days he is no longer infectious, but has recovered. So that is one less person infected. If he has infected more than one person by that time to 'compensate' for his own recovery, then the epidemic will grow. If on average he has managed to infect less than one person, then the epidemic is on the wane.

Yet another way to grasp the same point is to think about the average interval between causing infections, called the *serial interval*. If an infected person on average takes a time T to bring about a new infection and if T is less than D, then the infected person, on average, will be able to bring about a new infection before they recover and cease being infectious. On the other hand, if T is greater than D, then on average an infected person will have recovered before they can infect someone else. We can relate T to things we have already seen. T is the reciprocal or inverse of β. If someone infects two people per day, then on average that is half a day between infections. If someone infects four people per day, then that's the same as six hours between infections. The break-even point is when $T = D$, i.e. $1/\beta = D$ which is the same as $\beta D = 1$ and so $\mathscr{R}_0 = 1$. The serial interval for COVID-19 is thought to be around 4.

We can see why epidemiologists focus their attention on $\mathscr{R}_0$. It is much the same thing as focussing one's attention on β. But the idea that a key question is whether $\mathscr{R}_0$ is greater or less than one is attractive and easy to grasp. Measles has a very high $\mathscr{R}_0$, over 12, indicating that it is one of the most infectious diseases. Smallpox has a lower $\mathscr{R}_0$ at less than 6. The basic reproduction number for COVID-19 is thought to be between 2 and 3. This is roughly in the same range as the 1918 'flu pandemic, and more than the seasonal 'flu, whose $\mathscr{R}_0$ is between 1 and 2. It isn't possible to be precise about $\mathscr{R}_0$ because it varies from place to place both for the social reasons mentioned and also because the biology of a disease may be affected by climate. Avian 'flu A(H7N9), which appeared in several outbreaks among humans in China in 2013 and subsequent years is roughly ten times more deadly than COVID-19. But because its $\mathscr{R}_0$ among humans is less than 1 it was never able to cause an epidemic. The repeated outbreaks are caused by its continued presence among poultry which forms a so-called 'reservoir' for the virus.

6 The exponential growth of an epidemic disease

We have seen that if $\mathscr{R}_0$ is greater than 1, the disease will spread in the population. Let us return to our case where β, the number of new infections caused by a single

infective in each day, is two thirds and D, the duration of the infectious period, is 3 days. We saw that $\mathscr{R}_0$ is 2. This means that our individual patient, between becoming infected and his recovery three days later, has infected two other people. So we have doubled the number of infectious individuals. After a further three days these two individuals have recovered, but in the meantime they have each infected two further people, meaning that after six days there are four infected individuals. Likewise after nine days there are eight infectives, after twelve days there are sixteen infectives and so on. After one month there will be 1,024 infectives, and after two months, there will be over one million infectives. This is the exponential growth of infectives that one sees in the early phase of an epidemic.

Let us think about a real case, measles. Measles has a basic reproduction number of between 12 and 18—let us work with the more conservative 12. It has a very high $\mathscr{R}_0$ because the chance, τ, that a susceptible person will get measles after being exposed to an infectious person is very high—about 90%. A measles patient is infectious for about twenty days. If their infectiousness were the same throughout that period, then after two months a single infection in a susceptible population would have led to almost 2,000 cases, and about 2 to 5 deaths. After another month there will be over 70,000 cases, which would typically mean between 70 and 200 deaths.

7 Herd immunity and vaccination

The estimate of the number of cases of measles spreading in a susceptible population just given may well be an underestimate. That is because measles is most infectious in the early, prodromal phase when symptoms are just beginning to occur. The simplification of our model doesn't take this into account as it stands. If we were to treat the duration of the infectious period as being just a week, while $\mathscr{R}_0$ is 12, then after two months there would in principle the number of infected individuals would be 430 million.

Clearly that is unrealistic. We can only use $\mathscr{R}_0$ to tell us directly how many new infections there will be when almost everyone an infectious person encounters is susceptible. But if the disease spreads unchecked, then in due course more and more people will have had the disease and will have recovered. These people will make up a larger and larger proportion of the people that any infected person meets. As a result it will become more difficult for an infected individual to infect many others, and eventually they will not be able to infect anyone else before they have recovered. In this circumstance, the epidemic will die out, even though there are still some susceptible individuals left in the population. This is the state of herd immunity.

Let us recall the waning rule for epidemics *not* in their early stages:

$$\mathscr{R}_0 < \frac{N}{S}$$

As the epidemic continues the number of susceptibles, S, will get smaller and so N/S will get larger and must eventually be larger than $\mathscr{R}_0$.

The waning rule is therefore very helpful in telling us what proportion of the population needs to be immune in order for herd immunity to arise. That is particularly useful in deciding what proportion of the population needs to be vaccinated in order for there to be herd immunity. Herd immunity is important because not everyone can be vaccinated. Many babies and young children are too young to be vaccinated.

Some individuals are allergic to some vaccines. These people are protected by herd immunity.

Vaccination is a means of becoming immune without having been infected with the disease. That makes no difference to the use of the model—we can treat them just like recovereds. Let's take all the non-susceptibles to be immune. That's both the recovereds and the infectives. If immunity is brought about by vaccination then there should not be many infectives (if any), so we can treat the whole population as either susceptible or immune. The proportion of people who are immune is therefore:

$$\frac{N-S}{N}$$

which is the same as:

$$1-\frac{S}{N}$$

We can re-write the waning rule as follows:

$$\frac{S}{N} < \frac{1}{\mathscr{R}_0}$$

which is:

$$1-\frac{S}{N} > 1-\frac{1}{\mathscr{R}_0}$$

The left hand side is the proportion of the population that is immune, as we just saw. And the right hand side is equal to $(\mathscr{R}_0-1)/\mathscr{R}_0$. So we can re-write the waning rule as saying that there is herd immunity when:

$$\text{The proportion of the population that is immune is greater than } \frac{\mathscr{R}_0-1}{\mathscr{R}_0}$$

The value $(\mathscr{R}_0-1)/\mathscr{R}_0$ is often referred to the threshold for herd immunity and is then expressed as a percentage.

Herd immunity, in simple terms, occurs when a sufficiently large proportion of the population are immune that it is unlikely that any infected person will encounter and pass on their infection to any remaining susceptibles. More precisely, that probability is low enough that on average an infected person will have recovered before they pass on the disease to someone else. The susceptibles are therefore protected by the immunes.

We can look at a real example. We saw that $\mathscr{R}_0$ for measles is between 12 and 18. Let is start by assuming that it is 12. In that case the threshold for herd immunity is $(12-1)/12$ which is 92%. If $\mathscr{R}_0$ is 18, then the threshold for herd immunity is $18-1/18$ which is over 94%. Given such a high threshold for herd immunity, a sustained and vigorous programme of vaccination is needed to maintain it. If one did not employ vaccination, could herd immunity to measles be achieved naturally? For that to happen over 90% of the population would have to have had measles. And since measles has a fatality rate of 1 to 2 per 1,000, that would imply tens of thousands of deaths. Furthermore, about a quarter of cases require hospitalization, which would be an huge burden on health services. Furthermore, even if herd immunity were achieved, it would only be temporary, as a new generation is born who are susceptible but not immune—the latter will eventually reduce the proportion who are immune below the threshold, allowing the disease to take hold again if introduced from outside the population or if there are pockets of the disease in the population from which it has not entirely disappeared.

COVID-19 is rather less infectious than measles. $\mathscr{R}_0$ for COVID-19 is between 2 and 3. If it is 2.5, then for herd immunity about 60% of the population would need to be immune, either by vaccination or by having contracted the disease. Could herd immunity be achieved without a vaccine? That is something I might discuss in another video. For now we can calculate the number of deaths this would imply, if the vulnerable were not given special protection but were just as likely to be infected as anyone else. The 60% herd immunity threshold means that in the UK, with a population of 67 million, 40 million would have to be infected. The proportion of the infected who die—the infection fatality rate—is thought to be 0.66%. So the number of deaths would be over 260,000.

8 Assumptions and limitations

Earlier I emphasized that the SIR model makes a number of assumptions. That means that the model as it stands is limited in its applicability to those cases where the assumptions are true or near enough to the truth. I'll briefly mention a couple of such assumptions and limitations, though there are other we could discuss.

The model makes the assumption of *homogeneous mixing*. This is the idea that susceptibles, infectives, and recovereds are found in the same proportions throughout the population. That clearly is often not going to be entirely true—there may be geographical differences and there may be differences in other kinds of social group in the way that they interact with each other and with other people that means that homogeneity does not hold. The failure of the homogeneous mixing assumption can lead to an overestimate of the future size of an epidemic. At the same time, it might mean that although we have reached herd immunity threshold on average across the country, there are places or groups among whom the threshold has not been reached and among whom, therefore, an outbreak is possible.

Above I mentioned that one assumption is that everyone who contracts the disease will recover. Whereas we know that many such diseases will result in death. The model can be used to predict mortality, since we know what proportion of the infectives will die rather than recover. But that is for most diseases not a limitation on the model, since the deaths are not large in proportion to the total population or even in relation to the recovereds. With the policies currently in place, it is thought that the current COVID-19 will probably cause in the region of 20,000 deaths in the UK, quite possibly rather more, and so millions world wide These are terrible numbers which will bring with them a great deal of human misery. But as a proportion of the population that is less than one in three thousand, which is not enough to make a difference to the functioning of the model. On the other hand, if we were modelling the Black Death, then this would not be an assumption we could reasonably make.

The SIR model does make another assumption that is more relevant. It assumes that those who recover stay in the recovered category. But we know that immunity can sometimes be lost, in which case people will become susceptible again. If that is the case then we have to make a significant modification to the SIR model. We have to make it an SIRS (Susceptible–Infectious–Recovered–Susceptible) model, which will have consequences for the mathematical operation of the model. We do not yet know the degree to which this applies to COVID-19.

Another way in which the model can be improved is by acknowledging that diseases have a *latency* period. This is the period between becoming infected but before being infectious. Individuals during the latency period are no longer suscep-

tible (like both infectives and recovereds) but cannot pass on the disease (unlike infectives). Making an adjustment for this gives as the SEIR (Susceptible–Exposed–Recovered–Susceptible). This is different from the idea that a disease has an *incubation* period. The latter is the interval between being infected and the first symptoms appearing. If the incubation period is longer than the latency period, then there will be a time before symptoms appear that the patient is infectious.