# <1>Chapter 10: Complexities of epidemic modeling

In the prior Chapter, we derived and simulated the most basic epidemic model. We assumed that people can be in only one of three states (susceptible, infected, or recovered), and that people mix homogeneously throughout the population. In this Chapter, we examine how the Kermack-McKendrick model can be extended to simulate a wide variety of complex diseases and circumstances, and be adapted to incorporate the complex ways that people contact each other. We additionally describe methods for simulating individual behavior in response to an epidemic.

## <2>Extending the standard Kermack-McKendrick model

Once we leave the context of the Kermack-McKendrick model, the calculation of $R_0$ unfortunately gets complicated, often too complicated to produce any simple formulas that are memorable enough to be recalled for day-to-day public health practice (although we note that the $R_0$ can still be calculated using mathematics that are beyond the level of this textbook, as detailed elsewhere [1]). So, we often have to resort to simulation to identify what effect a disease will have in a population, and to measure the potential impact of a public health intervention on the disease.

We can extend our basic Kermack-McKendrick model simulation to incorporate a wide variety of diseases and interventions. Suppose, for example, that we want to model two potential interventions to reduce the burden of tuberculosis (TB). People who are infected with TB are not always infectious. Rather, people go into a “dormant" or “latent" state of disease (which is not dangerous, and not infectious), sometimes for several decades, before “activating" to the form of

active lung disease that we commonly associate with TB (and which is both infectious and potentially fatal if untreated).

As a public health planner, you want to compare two different strategies for reducing the burden of TB: (1) investment in a program to find and treat people with latent TB, preventing them from transitioning to active TB sometime in the future; and (2) investment in a program to expand treatment for people with active TB, preventing them from infecting susceptible people.

There is no way to compare these programs using the standard Kermack-McKendrick model, but we can extend the model to address the unique pathogenesis of TB. We can approach the problem with the same strategy we have used for all of our models: first draw the model diagram, then write down equations describing the diagram.

Our model diagram is shown in Figure 10.1. In the diagram, we have extended the classical Kermack-McKendrick model to account for the unique nature of TB. Here, we have extended the Kermack-McKendrick model to an "SEIR model", which is extends the SIR structure by adding a group of "exposed" people between the susceptible and infectious group. The exposed group is latently infected, but not yet infectious. They can move on to the infectious state at some rate $\gamma$. Also, since TB is a potentially deadly disease, suppose that there is a rate of death $\kappa$ from the active, infectious state.

[INSERT FIGURE 10.1 HERE]

We can write down a series of equations based on our model diagram:

[Equation 10.1] $$\frac{dS(t)}{dt} = \mu N(t) - \beta I(t)S(t) - \mu S(t)$$

[Equation 10.2] $$\frac{dE(t)}{dt} = \beta I(t)S(t) - \gamma E(t) - \mu E(t)$$

[Equation 10.3] $$\frac{dI(t)}{dt} = \gamma E(t) - vI(t) - \kappa I(t) - \mu I(t)$$

[Equation 10.4] $$\frac{dR(t)}{dt} = vI(t) - \mu R(t)$$

There are several differences between the SEIR model and the Kermack-McKendrick SIR model. First, we see that we have added one equation for the exposed (latent) TB state. People enter that state from the susceptible class, and flow to the active TB state at the rate $\gamma$. We have additionally added a rate of death from TB of $\kappa$ from the active, infectious state. Hence, we should not expect to have a constant population size in our model.

The textbook website has the *R* code corresponding to this model. Because the course of TB epidemics takes a long time to simulate, it would be painful to simulate in Excel (requiring hundreds of rows of equations), hence we will implement the code in *R* for ease. The code utilizes the following set of parameters: $N$ = 100,000 people, $\beta$ = 0.001, $\mu$ = 1/75 years$^{-1}$, $\gamma$ = 0.05 years$^{-1}$, $v$ = 2 years$^{-1}$, and $\kappa$ = 0.1 years$^{-1}$ (note that we are using convenient round numbers here and elsewhere in the book for teaching purposes; these parameters should not be misconstrued to reflect actual TB epidemics). We also set the initial conditions such that S = 99,999 people, E = 0 people, I = 1 person, and R = 0 people.

As in Chapter 9, we begin our *R* model code by first specifying our parameters:

```
N = 100000
beta = 0.001
mu = 1/75
gamma = 0.05
v = 2
kappa = 0.1
```

Second, we declare our time horizon of interest, let's say 20 years, and a small time step:

```
time = 20
dt = 0.01
```

Third, we insert our initial conditions and initiate some vectors for each state:

```
S = 99999
E = 0
I = 1
```

```
R = 0

Svec = S
Evec = E
Ivec = I
Rvec = R
```

Fourth, we insert our equations in two "for loops" (the first spanning across all times, and the second loop for each small time step within each time period). The current state in each loop is then added to the vectors for each state:

```
for (i in 1:time){
  for (i in 1:(1/dt)){
    S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
    E = E + beta*S*I*dt - gamma*E*dt - mu*E*dt
    I = I + gamma*E*dt - v*I*dt - kappa*I*dt - mu*I*dt
    R = R + v*I*dt - mu*R*dt
    Svec = c(Svec, S)
    Evec = c(Evec, E)
    Ivec = c(Ivec, I)
    Rvec = c(Rvec, R)
    N = S+E+I+R
  }
}
```

Note that we have updated the total population size *N* with each time step, since we are no longer assuming a constant population size because TB is deadly.

Finally, we can plot our results, adding the exposed group to our plot in purple, as shown in Figure 10.2:

```
plot(Svec,col="blue",type="l",xlab="time
steps",ylab="Pop",ylim=c(0,100000))
lines(Evec,col="purple")
lines(Ivec,col="red")
lines(Rvec,col="green")
legend(1500,100000,c("S","E","I","R"),lty=c(1,1,1,1),col=c("blue","purple"
,"red","green"))
```

As with our Kermack-McKendrick model in Chapter 9, we first plot one vector, specifying its color, line type, labels, and in the limits of the y-axis (from 0 to 100,000 people in

this case). Then we plot other vectors as lines, and add a legend, specifying the legend's position, text, line types, and colors; the commands produce Figure 10.2.

[INSERT FIGURE 10.2 HERE]

We see in Figure 10.2 that the curves for the disease are very different from the original Kermack-McKendrick model. The disease spreads over a longer period of time. Many people become latently infected, though only a few are actively infectious. Because Figure 10.2 is zoomed out, we can't see the infected line very clearly. Hence, we can just plot the infectious group by themselves:

```
plot(Ivec,col="red",type="l",xlab="time steps",ylab="Pop")
```

This produces Figure 10.3, where we see an interesting shape to the infectious pool, which peaks at a prevalence of about 2,000 people, with a very gradual rise and an even slower fall:

```
> max(Ivec)
[1] 2019.316
```

[INSERT FIGURE 10.3 HERE]

We can compare the two interventions we hope to contrast: (1) investment in a program to find and treat people with latent TB, preventing them from transitioning to active TB sometime in the future; and (2) investment in a program to expand treatment for people with active TB, preventing them from infecting susceptible people. Let's suppose that investment in treating people with latent TB would add a new rate that would remove people from the exposed state and put them in the recovered state at a rate $\tau = 0.1$ year$^{-1}$. Let's say that, by contrast, more investment in expanding treatment would increase the recovery rate to a value of 5 year$^{-1}$. Which program would make more impact on the total number of people who die from tuberculosis over the next 20 years?

For the treatment of latent TB, we can modify the equations slightly to add on the rate $\tau$ = 0.1 year$^{-1}$ shifting people from the latent to the recovered state. We first add the code:

```
tau = 0.1
```

We can include the code defining tau with the other parameters at the beginning of the code. We can then modify the equations to reflect the new rate of flow from the exposed to the recovered state:

```
S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
E = E + beta*S*I*dt - gamma*E*dt - mu*E*dt - tau*E*dt
I = I + gamma*E*dt - v*I*dt - kappa*I*dt - mu*I*dt
R = R + v*I*dt - mu*R*dt + tau*E*dt
```

We next need to determine what the overall number of people who die from TB would be over the time period, which is 20 years. We can add one more vector before the "for loop" to hold this quantity:

```
Deaths = 0
```

Then we can update its value within the "for loop":

```
Deaths = c(Deaths, kappa*I*dt)
```

When we run the model, we see that the total deaths after introducing the latent TB treatment program is about 1,400:

```
> sum(Deaths)
[1] 1398.232
```

Next, we can compare the number of deaths to the situation if we increased the recovery rate to a value of 3 year$^{-1}$. We first set the value of $\tau$ to 0, to show the absence of the latent treatment case, then set the value of $v$ to 5. Running the model in this scenario, we see that the total number of deaths has been dramatically reduced, to just under 700:

```
> sum(Deaths)
[1] 681.6078
```

The comparison between the two treatment alternatives can be visualized in Figure 10.4.

[INSERT FIGURE 10.4 HERE]

Our example reveals that we can extend the Kermack-McKendrick model in several ways to accommodate a variety of diseases and interventions. In addition to adding new states, we can add new flows between states. We can quickly extend the model in *R* and define the metrics we want to compare. In the next section, we'll also look at how to relax a key assumption we've used so far: The Law of Mass Action, which assumed that people are homogeneously mixing together and thereby all share roughly the same risk.

## <2>Heterogeneous mixing

Up to this point, we've been assuming *homogeneous mixing* of a population—meaning that people contact one another at a relatively even rate throughout the population, such that every person experiences the same average rate of disease. That's often a pretty good approximation to understand the general trends in large populations, but in some cases—for example, in the case of many sexually-transmitted diseases (STDs)—we need to modify that assumption. In the case of STDs, some people who have minimal amounts of sexual contact will be at very low risk for disease, while others who change sexual partners frequently will be at much higher risk. Many other diseases that present a major public health burden also involve such *heterogeneous mixing*, such as diseases that travel via injection drug use, or diseases that are segregated to only the poorest or most crowded areas of cities. Public health workers take a great deal of care to find the target population for their interventions, which often means determining those people who are at highest risk for disease, and providing both preventative and therapeutic interventions to them.

It has been observed that the American population tends to cluster around sexual activity “classes", in which a majority of people had a low level of sexual activity, and a smaller pool of people have very high rates of activity. A question has been whether it is more worthwhile to focus many resources on a few people at the high end of the activity distribution, or whether a more even spread of resources across more people would result in broader benefits at the population-level.

To address the dilemma of whether targeting treatment or broadening treatment might have more impact, let’s build a model to simulate a simple heterosexually-transmitted STD. Suppose we want a model of a disease like gonorrhea, in which infected/infectious people can be treated and return to the susceptible state, meaning that instead of an SIR model, we just have an SI model.

As with any other model, we can start the modeling process by drawing a diagram describing our model, followed by a set of equations translating our diagram into symbols.

[INSERT FIGURE 10.5 HERE]

In Figure 10.5, we have drawn an SI model, but created four subsets of persons for each of the four subsets of a heterosexual STD epidemic: high-activity men, high-activity women, low-activity men, and low-activity women. We end up with eight states to draw, because we have two states of disease, for each of two sexes, for each of two classes of sexual activity.

Since we are only simulating heterosexual transmission of the STD, we will have the rate of infection among men depend on the prevalence of infected women, and the rate of infection among women depend on the prevalence of infected men.

But a key challenge is how we can model the transmission rate. We need some strategy to overcome the Law of Mass Action. Rather than simply providing an estimate of the transmission

risk coefficient $\beta$ for all people, we need to account for the differences in risk levels between the high-activity and low-activity classes.

One strategy to account for the complex mixing of people would be to break up the general transmission risk coefficient $\beta$ into three component parts: (1) a part that has to do with the rate of risky contact (e.g., unprotected sexual encounters per year); (2) a second part that has to do with the biology of the organism being analyzed (e.g., the probability of transmission of the organism from one person to another, per unprotected sexual encounter between a susceptible and an infected person); and (3) a third part, which captures the probability of mixing with a person of the same or different activity level (e.g., is transmission from a high-activity or low-activity person). Let's say that the rate of unprotected contact is variable *c*, the probability of transmission of the organism from an infected to a susceptible person per unprotected contact is *p*, and the probability of mixing with a person of the same or different class is *w*. Hence, $\beta = c*p*w$. An advantage of breaking up one variable into two components, in this case, is that the interventions that we'll model may alter the rate of risky contact (e.g., educational interventions), and/or the probability of transmission (e.g., condom use), hence the extra complexity is justified since we know what the impact of STD-related interventions are on these parameters, but we don't know their effect on a generic transmission rate $\beta$.

Furthermore, it's not simply the case that highly-active people interact with people who have high activity, and vice versa. Otherwise we would have two different populations of people altogether; to the contrary, there are a proportion of people who mix between the two groups. This can be represented by a matrix, often called the "contact matrix", "mixing matrix", or "who acquires infection from whom" (WAIFW) matrix. Table 10.1 provides an example of a WAIFW matrix, which we note is not symmetrical because from the perspective of a high-activity person,

the chances of encountering another highly-active person is much greater than the probability of encountering a low-activity person.

[INSERT TABLE 10.1HERE]

To refer to this matrix, we can use conventional notation: w[row, column], such that w[1,1] means the probability a low activity person has unprotected sex with a low activity person, 0.6. Similarly, w[1,2] means the probability a low activity person has sex with a high activity person, 0.4. If the matrix were different between men and women, we could use two matrices (one for each gender), but we can assume the same matrix for men and women for simplicity in this example.

Next, we can write down equations for our model. Looking at our model diagram, we will have eight equations for each of the eight categories of people, where we can use the subscript *m* to refer to men, *f* to refer to women, *h* to refer to high-activity, and *j* to refer to low-activity. For simplicity, we will also assume that birth and death are such small rates from the perspective of a short-term STD epidemic that we will omit those rates from these equations and just focus on infection and recovery (rate *v*):

[Equation 10.5] $$\frac{dS_{m,h}}{dt} = -pc_h(w[2,1]I_{f,j} + w[2,2]I_{f,h})S_{m,h} + vI_{m,h}$$

[Equation 10.6] $$\frac{dI_{m,h}}{dt} = +pc_h\big(w[2,1]I_{f,j} + w[2,2]I_{f,h}\big)S_{m,h} - vI_{m,h}$$

[Equation 10.7] $$\frac{dS_{f,h}}{dt} = -pc_h(w[2,1]I_{m,j} + w[2,2]I_{m,h})S_{f,h} + vI_{f,h}$$

[Equation 10.8] $$\frac{dI_{f,h}}{dt} = +pc_h\big(w[2,1]I_{m,j} + w[2,2]I_{m,h}\big)S_{f,h} - vI_{f,h}$$

[Equation 10.9] $$\frac{dS_{m,j}}{dt} = -pc_j(w[1,1]I_{f,j} + w[1,2]I_{f,h})S_{m,j} + vI_{m,j}$$

[Equation 10.10] $$\frac{dI_{m,j}}{dt} = +pc_j\big(w[1,1]I_{f,j} + w[1,2]I_{f,h}\big)S_{m,j} - vI_{m,j}$$

[Equation 10.11] $$\frac{dS_{f,j}}{dt} = -pc_j(w[1,1]I_{m,j} + w[1,2]I_{m,h})S_{f,j} + vI_{f,j}$$

[Equation 10.12] $$\frac{dI_{f,j}}{dt} = +pc_j\left(w[1,1]I_{m,j} + w[1,2]I_{m,h}\right)S_{f,j} - vI_{f,j}$$

In each equation, we see that the infection rate is a product of the biological probability of transmission upon an unprotected encounter, the rate of such encounters (conditional on the activity class of the person), the probability that the infected person being contacted is of one class or another, and the prevalence of infection among such persons. The textbook website has accompanying *R* code to implement this model. As with prior models, we begin the model with the input parameters, which in this case include a contact vector with two components (a value of 5 per year for high-risk people and 1 per year for low-risk people), and a WAIFW matrix:

```
p = 0.00015
c = c(1,5)
w = matrix(c(.6,.4,.1,.9),ncol=2,byrow=TRUE)
v=2
```

We next set the time horizon for the model and the time steps:

```
time = 10
dt = 0.01
```

As with other models, we specify initial conditions and starting vectors for our states; in this case, we have arbitrarily split the population into 50% males and 50% females, with 10% of each sex in the high-activity class, and use a for loop to make it easier for us to keep track of the susceptible and infected people. In particular, we create a matrix for susceptible and for infected people, with rows corresponding to sexes and columns corresponding to activity classes (low or high activity class(:

```
S = matrix(ncol=2,nrow=2)
I = matrix(ncol=2,nrow=2)


for (sex in 1:2){
  for (group in 1:2){
    S[sex,group]=4000
```

```
          I[sex,group]=1000
        }
      }
```

Inside our "for loop", we insert the equations just as they appear in Equations 10.5 through 10.12, but using *R*'s ability to lookup the *c* vector and *w* matrix to find the right beta for each group of people and to fill in the values over a loop across both sexes and both activity classes, saving us from having to type out eight equations. Here we define a vector 'Numinf' to keep track of the total number of people infected in each time step:

```
beta=matrix(ncol=2,nrow=2)
Numinf = sum(I)
for (i in 1:(time/dt)){
      for (sex in 1:2){
          for (group in 1:2){
beta[sex,group] = p*c[group]*(w[group,1]*I[3-sex,1]+w[group,2]*I[3-sex,2])
S[sex,group] = S[sex,group] - beta[sex,group]*S[sex,group]*dt  +
v*I[sex,group]*dt
I[sex,group] = I[sex,group] + beta[sex,group]*S[sex,group]*dt -
v*I[sex,group]*dt
          }
        }
        Numinf = c(Numinf,sum(I))
}
```

Finally, after our "for loop", we can plot the total number of people infected over each time step:

```
plot(Numinf)
```

[INSERT FIGURE 10.6 HERE]

As shown in Figure 10.6, the number of infected people decreases then increases over time.

Overall, the example of the STD epidemic shows us how an SIR model can be extended when we wish to depart from the Law of Mass Action and instead take into account a heterogeneous mixing matrix. As illustrated in this example, even a simple deviation to two classes of risk produces a large number of equations and complexities for modeling. An even more complex strategy for modeling heterogeneous transmission is the approach of simulating dynamic social networks and how they influence disease transmission. Although the mathematics of network simulation are beyond the scope of this book, the 'EpiModel' package in *R* provides a user-friendly approach to simulate epidemics over networks, and can be installed and utilized just like any other package to produce simulations of epidemics incorporating dynamic social networks.

### <2>Vector-borne disease

Particularly in the context of climate change and rising temperatures that spread mosquitoes and other insects, infectious disease epidemiologists are increasingly concerned about the potential threat of increasing vector-borne (e.g., mosquito-driven) epidemics. When more than one species is involved in the spread of disease, it may not be obvious how we can adapt the Kermack-McKendrick model to address epidemics of vector-borne disease. Here, we illustrate how SIR-type models can be extended to address diseases where more than just the human species is involved.

Suppose, for example, that we wish to simulate the prevalence of malaria. We can simulate malaria as an SI model in which both humans and mosquitoes can be infected, as shown in Figure 10.7. Note that while humans can recover from infection, mosquitoes typically do not.

[INSERT FIGURE 10.7 HERE]

We can designate humans with subscript *h* and mosquitoes as subscript *m*. Humans catch malaria from mosquitoes, and mosquitoes catch malaria from humans. Translating these dynamics into equations, we have Equations 10.13 through 10.16:

[Equation 10.13] $\frac{dS_h}{dt} = \mu_h N_h + vI_h - \beta I_m S_h - \mu_h S_h$

[Equation 10.14] $\frac{dI_h}{dt} = \beta I_m S_h - vI_h - \kappa_h I_h - \mu_h I_h$

[Equation 10.15] $\frac{dS_m}{dt} = \mu_m N_m - \beta I_h S_m - \mu_m S_m$

[Equation 10.16] $\frac{dI_m}{dt} = \beta I_h S_m - \kappa_m I_m - \mu_m I_m$

We can see that the first two equations describe an SI model among humans with recovery rate $v$ and death rate from malaria $\kappa$. The rate of infection depends on the number of infected mosquitoes. Similarly, mosquitoes face an infection rate dependent on the number of infected humans. The *R* code corresponding to this model can be downloaded from the textbook website. First, we used the following parameters vectors, where the first element of each vector is the parameter corresponding to humans, and the second is the parameter corresponding to mosquitoes, with time units in years:

```
mu = c(.015,12)
beta = 0.00003
kappa = c(1/5, 10)
v = 2
```

Second, we specify our time horizon and time step:

```
time = 1
```

```
dt = 0.0001
```

Here, we've intentionally chosen to simulate just one year in very small time steps to illustrate the complex dynamics between humans and mosquitoes.

Next, we specify our initial conditions and populate empty vectors:

```
Sh = 90000
Ih = 10000
Sm = 900000
Im = 100000

Shvec = Sh
Ihvec = Ih
Smvec = Sm
Imvec = Im
```

Finally, we insert our equations into a "for loop":

```
for (i in 1:time){
  for (i in 1:(1/dt)){
    Nh = Sh+Ih
    Nm = Sm+Im
    Sh = Sh + mu[1]*Nh*dt + v*Ih*dt - beta*Im*Sh*dt - mu[1]*Sh*dt
    Ih = Ih + beta*Im*Sh*dt - v*Ih*dt- kappa[1]*Ih*dt - mu[1]*Ih*dt
    Sm = Sm + mu[2]*Nm*dt - beta*Ih*Sm*dt - mu[2]*Sm*dt
    Im = Im + beta*Ih*Sm*dt - kappa[2]*Im*dt - mu[2]*Im*dt
    Shvec = c(Shvec,Sh)
    Ihvec = c(Ihvec,Ih)
    Smvec = c(Smvec,Sm)
    Imvec = c(Imvec,Im)
  }
}
```

When we insert commands to plot our output, we obtain Figure 10.8:

```
plot(Shvec,lty=1,type="l",xlab="time steps",ylab="Pop",ylim=c(0,1000000))
lines(Ihvec,lty=2)
lines(Smvec,lty=3)
lines(Imvec,lty=4)
legend(7000,1000000,c("Sh","Ih","Sm","Im"),lty=c(1,2,3,4))
```

[INSERT FIGURE 10.8 HERE]

We see that because humans and mosquitoes live on such different time scales, the prevalence of malaria tends to change wildly as the two species infect each other. One can further introduce seasonality by adjusting the birth rate of mosquitoes to be a sine function over

time, enabling us to re-introduce more susceptible mosquitoes over time, in turn producing new waves of infection and a harmonic back-and-forth of infectious seasons.

Overall, we can see from this Chapter that the Kermack-McKendrick model is not limited in its capacity but quite versatile in how it can be extended to a wide range of infectious diseases and interventions, producing a complex and meaningful tool for public health planners to understand the complex dynamics of infectious diseases.