Captions

Introduction

Chapter 1

Figure 1.1

The bathtub analogy for deciphering incidence, prevalence, and mortality. We can conceptualize the *flow* of water into the bathtub (through the faucet) as the *incidence* rate, or the number of people over a period of time (per week, per month, or per year, to give a few examples) who newly acquire the disease. We can conceptualize the *level* of water in the bathtub at the *prevalence* of disease, or the number of people at a particular time of a study who have the disease. Unlike a rate of incidence, the prevalence of a disease reflects a single point in time—not a change in people over a period of time. We can conceptualize the *flow* of water out of the bathtub (through the drain) as the *mortality* rate from the disease, or the number of people over a period of time who have died from the disease.

Chapter 2

Figure 2.1

A decision tree for whether to recommend the new, experimental drug to a patient with kidney cancer, which specifies the probability of each outcome. If a person undergoes treatment for the

cancer with a standard, existing drug, they have a 15% chance of survival and an 85% chance of death. With the new, experimental drug, they have a 40% chance of remission and a 60% chance of death. However, among those people who experienced a remission, 50% survived and 50% later died. We typically designate the start of a decision tree with a square, and each "chance node" or decision branching point with a circle. Triangles then designate the endpoint (sometimes called a leaf node or terminal node) of each branch. Next to each decision branch is the probability of traveling along that branch. Next to each node is the value of the node.

Figure 2.2

Solution to the decision tree for whether to recommend the new, experimental drug to a patient with kidney cancer.

Figure 2.3

Decision tree for switching some low complexity patients to 'virtual visits'. The top half of the tree refers to the status quo situation, and the bottom half to the situation after virtual visits are initiated. Note that of the 30% of time spent on low complexity visits in the status quo situation, one-third (10%) are converted to virtual visits, leaving 20% of time for in-person visits. However, because virtual visits take half as much time as in-person visits, the time for low complexity virtual visits is only 5% of time. The other 5% is distributed (2.5% each) between moderate and high complexity visits.

Figure 2.4

Conceptualization of a quality-adjusted life-year (QALY), using tuberculosis treatment as an example. The standard treatment is six months and puts people into a moderate state of quality of life (0.8) because of its side-effects. The newer drug that modifies the treatment regimen makes the overall treatment course shorter (four months) but puts people into a lower state of quality of life (0.75) during that period, because of more severe side-effects.

Figure 2.5

A cost-effectiveness plane. The x-axis quantifies how effective a treatment is, in quality-adjusted life years (QALYs) or an equivalent metric. The y-axis quantifies how costly the treatment is, in dollars.

Chapter 3

Figure 3.1

The first stage of constructing a decision tree for the male pattern baldness problem, showing how the tree would extend infinitely long as people can perpetually go back to the start of the tree (if returning to a hairy state), or continue the tree (if in the frontal bossing or combover state).

Figure 3.2

State diagram for the Markov model of the male pattern baldness “epidemic”. There are four possible states: being hairy, having frontal bossing, having a combover, or being fully bald (a

"bowling ball"). The four health states are connected by state transition probabilities that describe how likely it is for a person to transition from one state to the other.

Figure 3.3

State diagram for the Markov model of the male pattern baldness "epidemic", including self-loops which reflect the probability of staying within each state from one time step to another.

Figure 3.4

State diagram for the Markov model of the male pattern baldness "epidemic", after introduction of an intervention that can shift people with a combover to the hairy state.

Chapter 4

Figure 4.1

Screenshot of *RStudio*. At the top left of the figure, we can see a white square with a "+" sign, which we should click to start a new "R Script". This opens a blank page for us to use. *RStudio* is organized into four sections by default, with a "Code" window that we just opened to type in commands; a "Console" window that appears in the top right of the figure (but maybe located on the bottom left corner of some systems), which tell us what the engine is doing and provides our output to our commands; a "Workspace" (bottom left of the figure, but occasionally top right on some systems) that shows us the names of our parameters and datasets, and a window with several tabs that allows us to find files, view plots, get help, and install packages.

Figure 4.2

Simple commands in R. In the code window (the blank page at the upper left that we created), we can type in 1+2, and then click the “Run” button at the top right of the window, which looks like a green arrow pointing out of a rectangle.

Figure 4.3

State diagram for the Markov model of the male pattern baldness “epidemic”.

Figure 4.4

Probability of a combover over time during the course of the simulation.

Figure 4.5

Histogram showing the distribution of incremental cost-effectiveness ($/QALY) for the simulated intervention.

Chapter 5

Figure 5.1

Peeps.

Figure 5.2

Illustration of a business model that involves two rates: a rate of purchasing product from the manufacturer, and a rate of selling product from your warehouse to the grocery store.

Figure 5.3

The classical Kermack-McKendrick “SIR” model depicting infectious disease epidemics in terms of susceptible, infectious, and recovered compartments.

Figure 5.4

The transmission of an infectious disease. In this example, every infected person produces three secondary infected people.

Figure 5.5

The transmission of an infectious disease after vaccination. Vaccination reduces the number of people susceptible to infection. Shown here is an example with one out of every three people in the population vaccinated. As shown, the number of secondary cases is reduced as vaccinated people interrupt the chain of transmission.

Figure 5.6

Graph of the SIR model in *R*. The book website contains a version of the code for download.

Figure 5.7

The histogram that tells us how much the size of our epidemic in the first year varied due to the uncertainty in our input parameters. The x-axis displays the value of ‘Totveciters’, or the total number of people infected over the course of the first year of the simulated epidemic across

multiple simulated iterations of the model. The y-axis displays the frequency of each value of Totveciters.

Chapter 6

Figure 6.1

Translation of the Kermack-McKendrick "SIR" model to the "SEIR" model by addition of the "exposed" state, which is not infectious but can progress to an infectious state.

Figure 6.2

Simulation of the SEIR model in *R*. Curves for the disease are very different from the original Kermack-McKendrick model. The x-axis reflects time steps (1/100 of the year) in the simulation, and the y-axis reflects the number of people in the model in each state. The variables "S", "E", "I", and "R" reflect each of the four states of disease shown in Figure 6.1.

Figure 6.3

Change in the number of infectious people over time in the SEIR model implemented in R. Note the y-axis range is changed from Figure 6.2.

Figure 6.4

Change to the numbers of deaths from tuberculosis based on the SEIR model, if we treat latently-infected persons or if we treat active infectious people.

Figure 6.5

The susceptible-infected model of a heterosexually-sexually-transmitted disease.

Figure 6.6

The number of infected people decreases then increases over time in the susceptible-infected model of a heterosexually-sexually-transmitted disease. The book website contains code in *R* to reproduce the graph.

Figure 6.7

Simulation of a malaria epidemic with susceptible and infectious humans and mosquitoes. While humans can recover from infection, mosquitoes typically do not.

Figure 6.8

Number of "susceptible humans" (Sh), "infected humans" (Ih), "susceptible mosquitoes" (Sm) and "infected mosquitoes" (Im) in a simulated malaria epidemic. The book website contains code in *R* to reproduce the graph.

Chapter 7

Figure 7.1

Decision tree for detecting severe acute malnutrition, without the prealbumin test.

Figure 7.2

Decision tree for detecting severe acute malnutrition, with the prealbumin test.

Figure 7.3

A Markov model of the malnutrition, without the treatment program.

Figure 7.4

A Markov model of the malnutrition, with the treatment program.

Figure 7.5

Probability of severe malnutrition over time through the simulation.

Figure 7.6

Modified SIR model to include waning immunity and death from infection.

Figure 7.7

Modified SIR model to add vaccination with waning vaccine-induced immunity.

Chapter 8

Figure 8.1

Markov model of type 2 diabetes.

Figure 8.2

Markov models of type 2 diabetes, in which each of the states are subdivided into different populations of people: those with a large waist circumference, those with high blood pressure, and those with high body mass index.

Figure 8.3

Scatter plot of pairs of values for the first 1,000 rows of each column against each other, in the microsimulation of type 2 diabetes screening. Pop[1:1000,1] refers to rows 1 through 1000, column 1 (the waist circumference). Similarly, column 2 corresponds to systolic blood pressure, column 3 to body mass index, and column 4 to fasting blood glucose. The x and y axes display values of each variable against each other.

Figure 8.4

We compare the performance statistics of survey instrument #1 by plotting a "receiver operating characteristic" (ROC) curve. The ROC curve plots 1 minus the specificity of an instrument against the sensitivity of an instrument, and a *C-statistic* is calculated as the area under the curve. Higher *C*-statistics (closer to 1) indicate a better ability of an instrument to discriminate people with the disease from those without.

Figure 8.5

Histogram of net revenues after hiring a new nurse at a clinic, using microsimulation. The x-axis shows the 'revresults' variable corresponding to net annual revenue at the clinic, and the y-axis shows the frequency of each value for net revenue across 100,000 repeated simulations. As

shown, most but not all simulations resulted in a positive net revenue after the new nurse was hired.

Figure 8.6

In an agent-based model of soda consumption, soda consumption becomes popular initially and then slows in popularity as the probability of consumption reaches a maximum consumption level.

Chapter 9

Figure 9.1

Classification and regression tree analysis (CART) showing how a history of diabetes ('hodm'), systolic blood pressure ('sbp') and prothrombin time ('pt') relate to the score for risk of stroke and bleeding from a new drug versus warfarin.

Figure 9.2

Example of overfitting. The complex model fits the data points perfectly, having a coefficient of determination (R^2) of 1, meaning that all of the variance in the outcome is explained by the model. However, this complex model would not be expected to correctly predicted values of y for x values obtained from a repeat of the experiment, or for x values higher or lower than the values shown.

Figure 9.3

Example of a simpler model that has a lower coefficient of determination (R^2), but would be expected to better capture the general relationship between x and y.

Figure 9.4

Illustration of the process of cross-validation from the training set.

Figure 9.5

Example of a calibration curve relating predicted outcome rate to observed outcome rate. A perfect model would have a curve linking these points that falls perfectly on the 45’ line.

Figure 9.6

Gradient boosting machines (GBM). The circles display the covariates (x variables) whose values determine each branch point, while the diamonds provide the tree-predicted probability of the outcome under study.

Figure 9.7

Random forest (RF). The circles display the covariates (x variables) whose values determine each branch point, while the diamonds provide the tree-predicted probability of the outcome under study.

Figure 9.8

Deep learning neural networks, which are based on a loose caricature of the brain as a series of neurons in which inputs (like features of patients in a dataset) are processed by a layer of

neurons, which then inform another layer of neurons, and so forth, until an output is achieved (e.g., risk of a disease event).

Figure 9.9

A variable importance plot showing the standardized coefficients from a learner (showing X variable #91 to be particularly important). The standardized coefficient corresponds to how much the outcome probability changes for a standard deviation change in the input variable.

Figure 9.10

A partial dependence plot showing the relationship between an individual variable (X variable #91 in the example code) and the predicted probability of an outcome (mean_response).

Chapter 10

Figure 10.1

Histogram of the distribution of per-patient costs per year from the collaborative care model (CCM) and the primary care behaviorist (PCB) model, based on 10,000 simulations.

Chapter 11

Figure 11.1

Two alternative models of human papillomavirus and cervical cancer. Pre-cancerous states are designated as cervical intraepithelial neoplasia (CIN) stages 1, 2, and 3.

Figure 11.2

An illustration of the danger of overfitting a model to data in a theoretical demonstration. We first generated data describing the prevalence of all cervical intraepithelial neoplasia (CIN) lesions over a 30-year period among a fictional cohort of young women. To do so, we used the more “realistic” (complex) model in Figure 11.1 and assigned typical parameter values for the rates of progression and regression between states (a 5% rate of progression to the next state and 50% rate of regression per year to the prior state), then added noise to the data by drawing randomly from a normal distribution with mean equal to average prevalence and standard deviation corresponding to the prevalence rate's standard deviation. We performed a common model “calibration” approach in which both the simple and complex model shown in Figure 11.1 were fitted to the first 20 years of the data (solid dots), starting from standard parameter uncertainty ranges for progression and regression of disease. Despite being the “real” model, the more complex model had numerous alternative parameter values fit the data, since there are so many uncertainties about the progression and regression rates that many combinations of parameters were able to produce reasonable fits. As shown, one of these fits (green) produced a pattern that poorly forecast future prevalence (hollow dots) despite fitting the earlier prevalence data (solid red dots). The more complex model (wavy line) actually has a better “fit” to the early prevalence data when judged by standard reduced chi-squared criteria than does the simpler model (less wavy line); but as illustrated here, it has substantially poorer performance in forecasting prevalence in future years. The more complex model did not perform poorly simply by chance; it did so because there was insufficient prior knowledge to inform the parameter values describing the process of progression and regression through pre-cancerous states, hence the model was susceptible to fitting too tightly to the noisy prevalence data (overfitting).

Figure 11.3

An illustration of the identifiability problem, using an example from HIV policy. Both a 1-month duration of acute infection with six secondary infections per month (top graph) and a 3-month duration of acute infection with two secondary infections per month (bottom graph) produce the same result of six infections per person during the acute infectious period. But the implications of the two different parameter sets are very different, as early treatment (dashed line) would be effective in preventing secondary infections only in the latter case.