1

Some History

The mathematical study of diseases and their dissemination is at most just over three centuries old. To give a full account of the history of the subject would require a book in itself. The interested reader may refer to Burnet and White (1972) for a natural history of diseases, to Fenner *et al.* (1988) for an account of smallpox and its eradication, and to Bailey (1975) and Anderson and May (1991) for an outline of the development of mathematical theories for the spread of epidemics. We shall be concerned with the more modest task of placing some of the recent epidemic models in perspective. We therefore present a selective account of historical highlights to illustrate the developments of the subject between the seventeenth and early twentieth centuries. Creighton (1894) gives a descriptive account of epidemics in Britain to the end of the nineteenth century, and Razzell (1977) for smallpox.

1.1 An empirical approach

The quantitative study of human diseases and deaths ensuing from them can be traced back to the book by John Graunt (b. 1620, d. 1674) *Natural and Political Observations made upon the Bills of Mortality* (1662). These *Bills* were weekly records of London parishes, listing the numbers and causes of deaths in the parishes. In his book, Graunt discussed various demographic problems of seventeenth century Britain. Four of his twelve chapters deal with the causes of death of individuals whose diseases were recorded in the *Bills*. These death records, kept irregularly from about 1592 onwards and continuously from 1603, provided the data on which Graunt based his observations.

Table 1.1. *Numbers of deaths due to eight causes, and related risks*

Causes	Deaths	Risk
1. Thrush, Convulsion, Rickets, Teeth and Worms; Abortives, Chrysomes, Infants, Liver-grown and overlaid	71,124	0.310
2. Chronical Diseases: Consumptions, Ague and Fever	68,271	0.298
3. Acute Diseases, and Miscellaneous	49,505	0.216
4. Plague	16,384	0.071
5. Small-pox, Swine-pox, Measles and Worms without Convulsions	12,210	0.053
6. Notorious Diseases: Apoplex, Gowt, Leprosy, Palsy, Stone and Strangury, Sodainly, etc.	5,547	0.024
7. Cancers, Fistulae, Sores, Ulcers, Impostume, Itch, King's Evil, Scal'd-head, Wens	3,320	0.014
8. Casualties: Drowned, Killed by Accidents, Murthured	2,889	0.013

Source: Graunt (1662). The figures for Groups 1, 2, 3, 6 and 8 are quoted directly in Graunt's text, while those for Groups 4, 5 and 7 are obtained from his complete table of casualties appended after his comments in 'The Conclusion'.

In the 20 years 1629–36 and 1647–58, there were 229 250 deaths recorded from 81 different causes. Table 1.1 consolidates these data into eight main groups. The relative risks of death from each of the eight causes are indicated in the column furthest to the right in Table 1.1.

The main killers were Groups 1, 2 and 3; Graunt was led to observe that

> whereas many persons live in great fear, and apprehension of some of the more formidable, and notorious diseases following [Group 6]; I shall onely [*sic*] set down how many died of each: that the respective numbers, being compared with the Total of 229,250, those persons may the better understand the hazards they are in.

The notorious diseases were further broken up by Graunt into the subcategories of Table 1.2. Among these, apoplexy appears to have been the largest killer. Graunt's analysis of the various causes of death provided the first systematic method for estimating the comparative risks of dying from the plague, as against the chronical or other diseases, for example.

These observations may well be considered to be the first approach to the theory of competing risks, a theory that is now well established among modern epidemiologists.

1.2 A deterministic model

A more theoretical approach to the effects of a disease, namely smallpox, was taken by Daniel Bernoulli (b. 1700, d. 1782) almost a century later. Smallpox was then widespread in many parts of Europe where it affected

Table 1.2. *Deaths due to notorious diseases*

Causes	Deaths	Risk $\times 10^{-3}$
1. Apoplex	1306	5.697
2. Cut of the stone	38	0.166
3. Falling sickness	74	0.323
4. Dead in the streets	243	1.060
5. Gowt	134	0.585
6. Head-ach	51	0.222
7. Jaundice	998	4.353
8. Lethargy	67	0.292
9. Leprosy	6	0.026
10. Lunatique	155	0.676
11. Overlaid, and starved	529	2.308
12. Palsy	423	1.845
13. Rupture	201	0.877
14. Stone and strangury	863	3.764
15. Sciatica	5	0.022
16. Sodainly	454	1.980
Total	5547	24.196

Source: Graunt (1662).

a large proportion of the population, being responsible for around 10% of the mortality of minors (cf. Bernoulli's model-based estimate in the last column of Table 1.3) while those who survived were immune to further attack but left scarred for life. In 1760 Bernoulli read his paper 'Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir' to the French Royal Academy of Sciences in Paris. His intention was to demonstrate that variolation, i.e. inoculation with live virus obtained directly from a patient with a mild case of smallpox, a procedure that usually conferred immunity, would reduce the death rate and increase the population of France. Bernoulli's argument is readily recognized as the following problem in competing risks.

Suppose first that a cohort of individuals born in a particular year has an age-specific *per capita* death rate $\mu(t)$ at age t. Then given an initial population size $\xi(0) \equiv \xi_0$, its size $\xi(t)$ at age t satisfies the equation

$$\dot{\xi}(t) = -\mu(t)\xi(t), \tag{1.2.1}$$

so

$$\xi(t) = \xi(0) \exp\big(-\int_0^t \mu(u)\,\mathrm{d}u\big) \equiv \xi(0)\mathrm{e}^{-M(t)} \equiv \zeta(t), \tag{1.2.2}$$

where $M(t)$ is the cumulative hazard. We shall use $\zeta(\cdot)$ below.

Consider another cohort subject to both the general *per capita* death rate $\mu(t)$ as above and a further hazard (infection) like smallpox with a constant infection rate β per individual per unit time. Individuals succumb

Table 1.3. *Age profile of population afflicted with smallpox (Bernoulli)*

Age (yrs) t	Age cohort Total $\xi_\beta(t)$	Age cohort Immune $z(t)$	Age cohort Suscept. $x(t)$	Smallpox Incidence	Smallpox Cumulative Deaths	Annual Mortality Total	Annual Mortality Smallpox
0	1,300	0	1,300				
1	1,000	104	896	137	17.1	300	17.1
2	855	170	685	99	29.5	145	12.4
3	798	227	571	78	39.2	57	9.7
4	760	275	485	66	47.5	38	8.3
5	732	316	416	56	54.5	28	7.0
6	710	351	359	48	60.5	22	6.0
7	692	381	311	42	65.7	18	5.2
8	680	408	272	36	70.2	12	4.5
9	670	433	237	32	74.2	10	4.0
10	661	453	208	28	77.7	9	3.5
11	653	471	182	24.4	80.7	8	3.0
12	646	486	160	21.4	83.4	7	2.7
13	640	500	140	18.7	85.7	6	2.3
14	634	511	123	16.6	87.8	6	2.1
15	628	520	108	14.4	89.6	6	1.8
16	622	528	94	12.6	91.2	6	1.6
17	616	533	83	11.0	92.6	6	1.4
18	610	538	72	9.7	93.8	6	1.2
19	604	541	63	8.4	94.8	6	1.0
20	598	542	56	7.4	95.7	6	0.9
21	592	543	48.5	6.5	96.5	6	0.8
22	586	543	42.5	5.6	97.2	6	0.7
23	579	542	37	5.0	97.8	7	0.6
24	572	540	32.4	4.4	98.3	7	0.5

Source: Bernoulli (1760). Note that Halley's table (column 1) starts at $t = 1$; Bernoulli gives reasons for choosing cohort size 1,300 for $t = 0$. Bernoulli used $\alpha = \beta = 1/8$, and obtained his figures by smoothing to the mid-point of the previous year, so his figure 17.1 for $t = 1$, coming from 1017.1, differs from $1014.9 = 8 \times 1000/[7 + \exp(-1/8)]$ as follows from (1.2.6) (cf. Gani, 1978).

only once, the result of such infection being either death in a fraction α of cases or immunity for the remainder of life in the complementary fraction $1 - \alpha$. Denote the number of individuals still susceptible to the disease at age t by $x(t)$, and the total number of the surviving cohort of age t, whether immune or not, by $\xi_\beta(t)$ as shown in Figure 1.1. To simplify the mathematical model, the infectious state is assumed to be instantaneous, so that as soon as an infection occurs, the infective individual either dies or recovers immediately. Then for the $x(t)$ susceptibles and $z(t) \equiv \xi_\beta(t) - x(t)$ immunes in this cohort,

$$\dot{x}(t) = -\big(\mu(t) + \beta\big)x(t) \tag{1.2.3a}$$

and

$$\dot{z}(t) = -\mu(t)z(t) + (1 - \alpha)\beta x(t). \tag{1.2.3b}$$

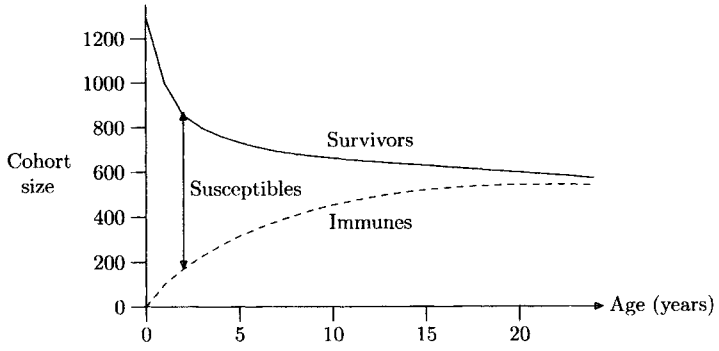


Figure 1.1. Survivors ξ_β (——) and immunes z (- - - -) in cohort of initial size 1300 (data from Table 1.3). The susceptibles of age t in the cohort are $x(t) = \xi_\beta(t) - z(t)$.

These equations are solved via integrating factors. Using $M(t)$ and $x(0) = \xi_\beta(0) = \xi(0) = \xi_0$ as before, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{M(t)+\beta t}x(t)\right) = 0,$$

whence

$$x(t) = \xi_0 \mathrm{e}^{-M(t)}\mathrm{e}^{-\beta t}, \tag{1.2.4}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{M(t)}z(t)\right) = (1-\alpha)\beta \mathrm{e}^{M(t)}x(t) = (1-\alpha)\beta\xi_0\mathrm{e}^{-\beta t}.$$

Integrating on $(0, t)$ and simplifying,

$$\begin{aligned}\xi_\beta(t) &= \mathrm{e}^{-M(t)}\xi_0[\mathrm{e}^{-\beta t} + (1-\alpha)(1-\mathrm{e}^{-\beta t})] \\ &= \zeta(t)(1-\alpha+\alpha\mathrm{e}^{-\beta t}),\end{aligned} \tag{1.2.5}$$

using (1.2.2); observe that $\zeta(t) = \xi_0(t)$ when the infection rate $\beta = 0$.

Equation (1.2.5) relates the sizes of the surviving cohorts of age t in populations with ($\beta > 0$) and without ($\beta = 0$) smallpox, respectively. Bernoulli used it in the form

$$\zeta(t) = \frac{\xi_\beta(t)}{1-\alpha+\alpha\mathrm{e}^{-\beta t}} \tag{1.2.6}$$

to estimate the size of a surviving cohort $\zeta(t)$ in a 'state without smallpox' on the basis of Halley's (1693) data from Breslau, now Wrocław. This estimation required parameters α and β; on reviewing what evidence he could

from a number of areas, Bernoulli fixed on $\alpha = \beta = 0.125$. Use of these estimates in (1.2.6) yields the data in the last column of Table 1.3; entries in the other columns are derived from this column and Halley's data. Observe that, granted the validity of Bernoulli's assumptions, smallpox caused between 10 and 40% of deaths between ages 2 and 23.

Suppose Bernoulli had had available observations of the form of his $\zeta(\cdot)$ at (1.2.6), for a 'state without smallpox' with a death rate similar to that $\mu(\cdot)$ prevailing in the areas from which Halley's data were drawn (column 2 of Table 1.3). Then taking differences of (1.2.5) with itself for times $t = t'$ and $t = t' + 1$ yields

$$\Delta\big(\xi_\beta(t')/\zeta(t')\big) \equiv \frac{\xi_\beta(t')}{\zeta(t')} - \frac{\xi_\beta(t'+1)}{\zeta(t'+1)} = \alpha(1 - \mathrm{e}^{-\beta})\mathrm{e}^{-\beta t'},$$

so that

$$\ln \Delta\big(\xi_\beta(t')/\zeta(t')\big) = -(a + \beta t'), \tag{1.2.7}$$

where $a = -\ln[\alpha(1 - \mathrm{e}^{-\beta})]$. This is the simplest way of expressing the result (1.2.5) for the purpose of estimating β and α *conditional* on such extended data being available. All that Bernoulli could do was to present the advantage of variolation (i.e. absence of deaths due to smallpox) on the basis of his model-based calculations. Note too that the population risk of death from smallpox (cf. Tables 1.1–2) as implied by Table 1.1 is about $100/1300 \approx 7.7\%$, higher than in Table 1.1 because the London population from which Graunt drew his data, included more immigrants than Breslau. In Halley's day Breslau had rather few immigrants, and hence, proportionately more infant and childhood deaths, smallpox being more prevalent amongst children than adults.

1.2.1 The Law of Mass Action

The *Law of Mass Action* has found wide applicability in many areas of science. In chemistry, the idea that a reaction is influenced by the quantities of the reactant materials goes back at least to Boyle (*c.* 1674). Around 1800, C. L. Berthollet emphasized the importance of mass or concentration of a substance on a chemical reaction, but this was not generally accepted for half a century. Ultimately, Guldberg and Waage (1864–1867) postulated that *for a homogeneous system, the rate of a chemical reaction is proportional to the active masses of the reacting substances* (Glasstone (1948, p. 816)).

Applied to population processes, *if the individuals in a population mix homogeneously, the rate of interaction between two different subsets of the*

population is proportional to the product of the numbers in each of the subsets concerned. In any population it is possible for several processes to occur concurrently, in which case the effects on the numbers in any given subset of the population from these various processes are assumed to be additive. Thus, in the case of epidemic modelling, the Law is applied to rates of transition of individuals between two interacting categories of the population, such as susceptibles who, as a result of contact with infectives, themselves become infectives; a second simultaneous process is that of the infectives who become removals. These two processes underlie equations (1.3.2a) and (1.3.2c) respectively: when more than one process is involved, as for the numbers of infectives in equation (1.3.2b), the effects are additive.

Application of the Law to transitions that occur in discrete time is not so straightforward, but, subject to certain constraints on the size of the change involved (see e.g. Section 2.8 below), it remains valid.

The Law also has a stochastic version when the process concerned is assumed to be Markovian, and the rate is then interpreted as the infinitesimal transition probability.

Implicit in the 'proportionality' aspect of the Law, is an assumption that the quantities concerned in inducing the transition are subject to *homogeneous mixing* with each other. The Law can then be seen as the result of superposing all possible contributions of the individual components to the interaction, these individuals being regarded as equally likely to interact with each other in a given (small) interval of time.

1.3 From curve-fitting to homogeneous mixing models

First issued in 1837, each *Annual Report of the Registrar-General of Births, Deaths and Marriages in England* included tables of causes of death and commentaries. The *Report* for 1840 includes a contribution from William Farr[1] entitled 'Progress of epidemics', in which Farr attempted to characterize mathematically the smoothed quarterly data for smallpox deaths. Some 26 years later, in a letter to the London *Daily News* of 17 February

[1]Farr was appointed compiler of abstracts to the General Register Office in 1839 and remained there until retirement in 1879. Early volumes of the *Annual Reports* contain papers of Farr prefaced by a 'Letter to the Registrar-General'; they cover a variety of issues pertaining to the data in the *Reports*. Thus, in the *Sixth Annual Report* (1842) Farr noted that the annual small-pox death-rates per 10^6 live individuals for the years 1838–42 were 1 101, 604, 679, 408 and 172 respectively, and remarked that 'The reduction in the mortality from small-pox since 1840 was probably the result, at least in part, of the Vaccination Act' [of 1840]. Later he gave the 1850 death-rate as 263.

Table 1.4. *Deaths from smallpox in consecutive quarters 1837–39*

	Sum. 1837	Aut. 1837	Win. 1838	Spr. 1838	Sum. 1838	Aut. 1838	Win. 1839	Spr. 1839	Sum. 1839	Aut. 1839
Observed deaths	2513	3289	4242	4484	3685	3851	2982	2505	1533	1730
Deaths averaged over two consecutive quarters		2901	3766	4365	4087	3767	3416	2743	2019	1637
Percentage change			+30	+16	−6	−8	−9	−20	−26	−19

Source: Farr (1840).

1866 (quoted by Brownlee, 1915), he attempted to predict the spread of rinderpest among cattle by a similar method.

Table 1.4 gives the observed deaths in the smallpox epidemic of 1837–39 drawn from Farr (1840), together with the average values of consecutive quarters, for 10 quarters in all. Farr concluded that as the epidemic declined, he could detect an approximately steady rate of deceleration in the number of deaths per quarter. Brownlee (1906) later carried out work of a similar type: he fitted Pearson curves to epidemic data for several diseases, and for several different locations.

But these pragmatic approaches were essentially limited, so long as there was not an appropriate theory to explain the mechanism by which epidemics spread. By the beginning of the twentieth century, the idea of passing on a bacterial disease through contact between susceptibles and infectives had become familiar, and Hamer (1906) first foreshadowed the simple 'mass action' principle for a deterministic epidemic model in discrete time. This principle, which incorporates the principle of homogeneous mixing, has been the basis of most subsequent developments in epidemic theory (see Section 1.2.1 above and Anderson and May (1991, p. 7) for discussion).

Hamer, noticing the rise and fall of infectives in the course of a large range of epidemics, argued against variable infectivity. Specifically, he wrote that to explain the eventual decline of an epidemic, 'the assumption of loss of virulence or infecting power on the part of the organism is quite unnecessary'. He also put forward a numerical argument about the initial increase and eventual decline of the number of infectives in a population; this indicates that he was aware that both susceptibles and infectives affected the number of new infectives listed in the weekly reports of measles in London:

> Now the outbreak will take much longer to decline to extinction than it took to rise, for those especially exposed have in large part been already attacked and the disease must spread, in the main, among persons whose manner of life brings them comparatively little into contact with their fellows.

Let x_t, y_t be the numbers of susceptibles and infectives respectively at times $t = 0, 1, 2, \ldots$. Hamer's idea was equivalent to expressing the new number of infectives at time $t+1$ by Δy_t such that

$$\Delta y_t = \beta x_t y_t \qquad (t = 0, 1, 2, \ldots) \tag{1.3.1}$$

where the constant β is such that $\beta x_t y_t \le x_t$ for all t, i.e. $\beta \le 1/(\max_{i \ge 1} y_i)$. These new infectives are a proportion β of the number $x_t y_t$ of contacts between susceptibles and infectives, where β is known as the infection parameter. Because of the constraint on β, it follows that in a closed population in which $y_t = N - x_t$, we have $\beta x_t(N - x_t) \le x_t$ or $\beta(N - x_t) \le 1$; this is certainly satisfied if $\beta \le 1/N$.

Continuous time versions of epidemic equations were used by Ross (1916) and Ross and Hudson (1917) in their studies of populations subject to infection. But the form of equations most commonly used to characterize the typical *general epidemic* with susceptibles $x(t)$, infectives $y(t)$ and immunes $z(t)$ (such as a measles epidemic) is due to Kermack and McKendrick (1927). They assumed a fixed population size $N = x(t) + y(t) + z(t)$, and using the homogeneous mixing principle for continuous time $t \ge 0$ derived the (now) classical equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\beta xy, \tag{1.3.2a}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \beta xy - \gamma y, \tag{1.3.2b}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \gamma y, \tag{1.3.2c}$$

subject to the initial conditions $\big(x(0), y(0), z(0)\big) = (x_0, y_0, 0)$. Here β is[2] the infection parameter, similar to that in (1.3.1), and γ is the removal parameter giving the rate at which infectives become immune. In cases where death or isolation may occur, $z(t)$ represents all removals from the population, including immunes, deaths and isolates.

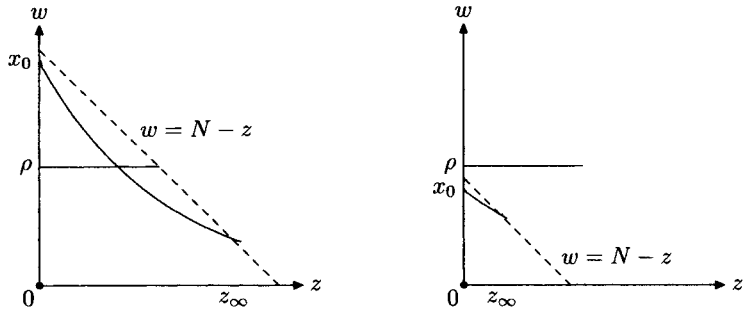
Dividing equation (1.3.2a) by (1.3.2c) gives

$$\frac{\mathrm{d}x}{\mathrm{d}z} = -\frac{\beta}{\gamma}x = -\frac{x}{\rho} \qquad \text{with} \qquad \rho \equiv \frac{\gamma}{\beta},$$

the parameter ρ being the *relative removal rate*. The solution of this equation is

$$x = x_0 \mathrm{e}^{-z/\rho},$$

[2]Some authors write $\beta = \beta'/x(0)$ so (1.3.2a) becomes $\dot{x} = -\beta'(x/x_0)y$.



(a) $x_0 < \rho$ (b) $x_0 > \rho$

Figure 1.2. z_∞ as the point of intersection of $w = N - z$ and $w = x_0\,\mathrm{e}^{-z/\rho}$.

so that

$$y = N - z - x_0 \mathrm{e}^{-z/\rho}.$$

Hence

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \gamma(N - z - x_0 \mathrm{e}^{-z/\rho}),$$

with the parametric solution

$$\gamma t = \int_0^z \frac{\mathrm{d}v}{N - v - x_0 \mathrm{e}^{-v/\rho}} \qquad (0 \le t < \infty). \tag{1.3.3}$$

Kermack and McKendrick obtained two basic results, referred to as their Threshold Theorem. The first is a criticality statement, and comes from equation (1.3.2b). Writing this equation as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \beta y(x - \rho)$$

shows that if the epidemic is ever to grow, then we must have $\mathrm{d}y/\mathrm{d}t\big|_{t=0} > 0$, or $x_0 > \rho$, i.e. *the initial number of susceptibles must exceed a threshold value* ρ. The second, deduced from (1.3.3), states that as $t \to \infty$, $z(t) \to z_\infty < N$. Now in the limit $t \to \infty$, $N - z_\infty - x_0 \mathrm{e}^{-z_\infty/\rho} = 0$ (see the graphs in Figure 1.2).

Suppose now that x_0 is close to N; then z_∞ will be the approximate solution of

$$0 = N - z_\infty - N\mathrm{e}^{-z_\infty/\rho} \approx N - z_\infty - N\Big(1 - \frac{z_\infty}{\rho} + \frac{z_\infty^2}{2\rho^2}\Big), \tag{1.3.4}$$