Bulletin of Mathematical Biology Vol. 53, No. 1/2, pp. 33–55, 1991.
Printed in Great Britain.

0092-8240/91$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

CONTRIBUTIONS TO THE MATHEMATICAL THEORY OF EPIDEMICS—I*

■ W. O. KERMACK and A. G. MCKENDRICK
Laboratory of the Royal College of Physicians,
Edinburgh, U.K.

1. Introduction. One of the most striking features in the study of epidemics is the difficulty of finding a causal factor which appears to be adequate to account for the magnitude of the frequent epidemics of disease which visit almost every population. It was with a view to obtaining more insight regarding the effects of the various factors which govern the spread of contagious epidemics that the present investigation was undertaken. Reference may here be made to the work of Ross and Hudson (1915–17) in which the same problem is attacked. The problem is here carried to a further stage, and it is considered from a point of view which is in one sense more general. The problem may be summarized as followed: One (or more) infected person is introduced into a community of individuals, more or less susceptible to the disease in question. The disease spreads from the affected to the unaffected by contact infection. Each infected person runs through the course of his sickness, and finally is removed from the number of those who are sick, by recovery or by death. The chances of recovery or death vary from day to day during the course of his illness. The chances that the affected may convey infection to the unaffected are likewise dependent upon the stage of the sickness. As the epidemic spreads, the number of unaffected members of the community becomes reduced. Since the course of an epidemic is short compared with the life of an individual, the population may be considered as remaining constant, except in as far as it is modified by deaths due to the epidemic disease itself. In the course of time the epidemic may come to an end. One of the most important problems in epidemiology is to ascertain whether this termination occurs only when no susceptible individuals are left, or whether the interplay of the various factors of infectivity, recovery and mortality, may result in termination, whilst many susceptible individuals are still present in the unaffected population.

It is difficult to treat this problem in its most general aspect. In the present

* Reprinted from the *Proceedings of the Royal Society*, Vol. 115A, pp. 700–721 (1927) with the permission of The Royal Society.

communication discussion will be limited to the case in which all members of the community are initially equally susceptible to the disease, and it will be further assumed that complete immunity is conferred by a single infection.

It will be shown in the sequel that with these reservations, the course of an epidemic is not necessarily terminated by the exhaustion of the susceptible members of the community. It will appear that for each particular set of infectivity, recovery and death rates, there exists a critical or threshold density of population. If the actual population density be equal to (or below) this threshold value the introduction of one (or more) infected person does not give rise to an epidemic, whereas if the population be only slightly more dense a small epidemic occurs. It will appear also that the size of the epidemic increases rapidly as the threshold density is exceeded, and in such a manner that the greater the population density at the beginning of the epidemic, the smaller will it be at the end of the epidemic. In such a case the epidemic continues to increase so long as the density of the unaffected population is greater than the threshold density, but when this critical point is approximately reached the epidemic begins to wane, and ultimately to die out. This point may be reached when only a small proportion of the susceptible members of the community have been affected.

Two of the reasons commonly put forward as accounting for the termination of an epidemic, are (1) that the susceptible individuals have all been removed, and (2) that during the course of the epidemic the virulence of the causative organism has gradually decreased. It would appear from the above results that neither of these inferences can be drawn, but that the termination of an epidemic may result from a particular relation between the population density, and the infectivity, recovery, and death rates.

Further, if one considers two populations identical in respect of their densities, their recovery and death rates, but differing in respect of their infectivity rates, it will appear that epidemics in the population with the higher infectivity rate may be great as compared with those in the population with the lower infectivity rate, especially if the density of the former population is in the neighbourhood of the threshold value. If, then, the density of a particular population is normally very close to its threshold density it will be comparatively free from epidemic, but if this state is upset, either by a slight increase in population density, or by a slight increase in the infectivity rate, a large epidemic may break out. Such great sensitiveness of the magnitude of the epidemic with respect to these two factors, may help to account for the apparently sporadic occurrence of large epidemics, from very little apparent cause. Further, it will appear that a similar state of affairs holds with respect to diseases which are transmitted through an intermediate host. In this case the product of the two population densities is the determining factor, and no epidemic can occur when the product falls below a certain threshold value.

2. General Theory. We shall first consider the equations which arise when the time is divided into a number of separate intervals, and infections are supposed to take place only at the instant of passing from one interval to the next, and not during the interval itself. We shall take the size of this interval, which at present may be considered constant, as the unit of time, and we shall denote the number of individuals in unit area at the time t who have been infected for θ intervals by $v_{t,\theta}$. The total number who are ill at this interval t is $\sum_{\theta=0}^{t} v_{t,\theta}$, which we shall call y_t. It should be noted that $v_{t,0}$ denotes the number of individuals at the time t who are at the beginning of their infection. Also we shall use the symbol v_t to denote the number who actually undergo the process of infection during the transition from the interval $t-1$, to the interval t. In general $v_{t,0}=v_t$ except at the origin, where we assume that a certain number y_0 of the population have just been infected, although this infection is naturally dependent on some process outside that defined by the equations which we shall develop. Thus:

$$v_{0,0}=v_0+y_0. \tag{1}$$

The whole process is indicated in the following schema:

Fresh infections	Numbers at each stage of illness	Number ill
v_3	$v_{3,0} \nearrow v_{3,1} \nearrow v_{3,2} \nearrow v_{3,3}$	y_3
v_2	$v_{2,0} \nearrow v_{2,1} \nearrow v_{2,2}$	y_2
v_1	$v_{1,0} \nearrow v_{1,1}$	y_1
v_0	$v_{0,0}$	y_0

The arrows indicate the course followed by each individual until he recovers or dies.

If ψ_θ denotes the rate of removal, that is to say it is the sum of the recovery and death rates, then the number who are removed from each θ group at the end of the interval t is $\psi_\theta v_{t,\theta}$, and this is clearly equal to $v_{t,\theta}-v_{t+1,\theta+1}$.

Thus:

$$\begin{aligned} v_{t,\theta} &= v_{t-1,\theta-1}(1-\psi(\theta-1)) \\ &= v_{t-2,\theta-2}(1-\psi(\theta-1)\,(1-\psi(\theta-2)) \\ &= v_{t-\theta,0}B_\theta, \end{aligned} \tag{2}$$

where B_θ is the product $(1-\psi(\theta-1))\,(1-\psi(\theta-2))\ldots(1-\psi(0))$.

Now v_t denotes the number of persons in unit area who became infected at the interval t, and this must be equal to $x_t \sum_1^t \phi_\theta v_{t,\theta}$ where x_t denotes the number of individuals still unaffected, and ϕ_θ is the rate of infectivity at age θ. (It is indifferent whether we include the term $\phi_0 v_{t,0}$ or not, since in this paper

we assume that ϕ_0 is zero, that is that an individual is not infective at the moment of infection.) This follows since the chance of an infection is proportional to the number of infected on the one hand, and to the number not yet infected on the other.

It is clear that:

$$x_t = N - \sum_0^t v_{t,0}$$

$$= N - \sum_0^t v_t - y_0, \tag{3}$$

where N is the initial population density.

If z_t denotes the number who have been removed by recovery and death, then:

$$x_t + y_t + z_t = N. \tag{4}$$

Thus we have:

$$v_t = x_t \sum_1^t \phi_\theta v_{t,\theta} = x_t \sum_1^t \phi_\theta B_\theta v_{t-\theta 0} \quad \text{(by 2)}$$

$$= x_t \left(\sum_1^t A_\theta v_{t-\theta} + A_t y_0 \right) \quad \text{(by 1)}, \tag{5}$$

where A_θ is written for $\phi_\phi B_\theta$.
Also:

$$y_t = \sum_0^t v_{t,\theta} = \sum_0^t B_t v_{t-\theta} + B_t y_0. \tag{6}$$

By definition:

$$-v_t = x_{t+1} - x_t, \tag{7}$$

hence equation (5) may be written:

$$x_t - x_{t+1} = x_t \left(\sum_1^t A_\theta v_{t-\theta} + A_t y_0 \right). \tag{8}$$

Also $z_{t+1} - z_t$ is the number of persons who are removed at the end of the interval of time t, and this is equal to $\sum_1^t \psi_\theta v_{t,\theta}$, i.e. to $\sum_1^t \psi_\theta B_\theta v_{t-\theta} + \psi_t B_t y_0$, hence writing C_θ for $\psi_\theta B_\theta$ we have:

$$z_{t+1}-z_t=\sum_1^t C_\theta v_{t-\theta}+C_t y_0. \tag{9}$$

Also by (4):

$$y_{t+1}-y_t=x_t\left[\sum_1^t A_\theta v_{t-\theta}+A_t y_0\right]-\left[\sum_1^t C_\theta v_{t-\theta}+C_t y_0\right]. \tag{10}$$

If now we allow the subdivisions of time to increase in number so that each interval becomes very small, then in the limit the above equations (4) and (7)–(9) become:

$$x_t+y_t+z_t=N. \tag{11}$$

$$v_t=-\frac{\mathrm{d}x_t}{\mathrm{d}t}, \tag{12}$$

$$\frac{\mathrm{d}x_t}{\mathrm{d}t}=-x_t\left[\int_0^t A_\theta v_{t-\theta}\,\mathrm{d}\theta+A_t y_0\right], \tag{13}$$

$$\frac{\mathrm{d}z_t}{\mathrm{d}t}=\int_0^t C_\theta v_{t-\theta}\,\mathrm{d}\theta+C_t y_0, \tag{14}$$

and from (6):

$$y_t=\int_0^t B_\theta v_{t-\theta}\,\mathrm{d}\theta+B_t y_0, \tag{15}$$

where:

$$B_\theta=\exp\left(-\int_0^\theta \psi(\alpha)\,\mathrm{d}\alpha\right), \qquad A_\theta=\phi_\theta B_\theta, \qquad \text{and} \qquad C_\theta=\psi_\theta B_\theta.$$

It can, however, be shown that these five relations are not independent and in fact that (11) is a necessary consequence of (13)–(15). The four independent relations (12)–(15) determine the four functions x, y, z and v.

By equation (13), dropping the suffix t except when necessary in the analysis,

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=-x\left[\int_0^t A_\theta v_{t-\theta}\,\mathrm{d}\theta+A_t y_0\right]\\&=-x\left[\int_0^t A_{t-\theta} v_\theta\,\mathrm{d}\theta+A_t y_0\right]\\&=x\left[\int_0^t A_{t-\theta}\frac{\mathrm{d}x_\theta}{\mathrm{d}\theta}\,\mathrm{d}\theta-A_t y_0\right],\end{aligned}$$

where x in the integral is now a function of θ.

Therefore:

$$\frac{\mathrm{d}\log x}{\mathrm{d}t} = A_{t-\theta}x_\theta\Big|_0^t - \int_0^t x_\theta \frac{\mathrm{d}A_{t-\theta}}{\mathrm{d}\theta}\,\mathrm{d}\theta - A_t y_0$$

$$= A_0 x_t - A_t x_0 + \int_0^t x_\theta A'_{t-\theta}\,\mathrm{d}\theta - A_t y_0,$$

where:

$$A'_{t-\theta} = \frac{\mathrm{d}A_{t-\theta}}{\mathrm{d}(t-\theta)} = -\frac{\mathrm{d}A_{t-\theta}}{\mathrm{d}\theta}.$$

But $A_0 = \phi_0 B_0 = \phi_0 = 0$, since we assume that an individual at the moment of becoming infected cannot transmit infection.

Hence:

$$\left.\begin{aligned}\frac{\mathrm{d}\log x}{\mathrm{d}t} &= -A_t(x_0+y_0) + \int_0^t x_\theta A'_{t-\theta}\,\mathrm{d}\theta \\ &= -A_t N + \int_0^t A'_\theta x_{t-\theta}\,\mathrm{d}\theta.\end{aligned}\right\} \tag{16}$$

We have not been able to solve this equation in such a way as to give x in terms of t as an explicit function. It may, however, be pointed out that this is an integral equation similar to Volterra's equation:

$$f(t) = \phi(t) + \int_0^t N(t,\theta)\phi(\theta)\,\mathrm{d}\theta,$$

except that in place of $f(t)$ we have $\mathrm{d}\log f(t)/\mathrm{d}t$.

If we consider an equation of the form:

$$\frac{\mathrm{d}\log x}{\mathrm{d}t} = A_t + \lambda\int_0^t N(t,\theta)x(\theta)\,\mathrm{d}\theta,$$

of which the above equation is a particular example, it would appear that a solution can be arrived at by a series of successive approximations in a way similar to the method used in resolving Volterra's equation.

We may write:

$$x = f_0(t) + \lambda f_1(t) + \lambda^2 f_2(t) + \cdots$$

It is easily seen that after substituting this expression in the equation:

$$\frac{dx}{dt} = x\left[A_t + \lambda \int_0^t N(t, \theta)x(\theta)\, d\theta\right],$$

and equating the coefficients of the powers of λ, we obtain:

$$\frac{d}{dt}f_n(t) = f_n(t)A_t + f_{n-1}(t)\int_0^t N(t, \theta)f_0(\theta)\, d\theta + f_{n-2}(t)\int_0^t N(t, \theta)f_1(\theta)\, d\theta$$
$$+ \cdots + f_0(t)\int_0^t N(t, \theta)f_{n-1}(\theta)\, d\theta$$
$$= L_{n-1}(t) \quad \text{say.}$$

This is a differential equation for $f_n(t)$ of which the solution is:

$$f_n(t)\exp\left(-\int_0^t A_t\, dt\right) = \int_0^t L_{n-1}(t)\exp\left(-\int_0^t A_t\, dt\right) dt + \text{constant},$$

where $L_{n-1}(t)$ is a function of the f's.

Also $f_n(0)$ is zero $(n>0)$, since the initial conditions are presumably independent of λ. Hence the constants of integration are all zero except $f_0(0)$.

In the case of this function we have:

$$\frac{df_0(t)}{dt} = f_0(t)A_t,$$

whence:

$$f_0(t) = f_0(0)\exp\left(\int_0^t A_t\, dt\right),$$

so that $f_0(0) = x_0$.

We thus have for the solution of the integral equation:

$$x = x_0E_t + \sum_{n=1}^{\infty} \lambda^n E_t \int_0^t \frac{L_{n-1}(t)}{E_t}\, dt$$
$$= E_t\left[x_0 + \sum_1^{\infty} \lambda^n \int_0^t \frac{L_{n-1}(t)}{E_t}\, dt\right],$$

where E_t is written for $\exp(\int_0^t A_t\, dt)$; and when $\lambda = 1$:

$$x = E_t\left[x_0 + \sum_1^{\infty} \int_0^t \frac{L_{n-1}(t)}{E_t}\, dt\right]. \tag{17}$$

Returning to equation (16) let us consider it in the rather more general form

$$\frac{\mathrm{d}\log x}{\mathrm{d}t} = A_t + \int_0^t Q_{t-\theta} x_\theta \, \mathrm{d}\theta.$$

Multiplying both sides by $\exp(-zt)$ where the real part of z is positive, and integrating with respect to t between the limits zero and infinity, we have:

$$\int_0^\infty \exp(-zt) \frac{\mathrm{d}\log x}{\mathrm{d}t} \, \mathrm{d}t = \int_0^\infty \exp(-zt) A_t \, \mathrm{d}t + \int_0^t Q_{t-\theta} x_\theta \, \mathrm{d}\theta \, \mathrm{d}t,$$

therefore:

$$-\log x_0 + \int_0^\infty z \exp(-zt) \log x \, \mathrm{d}t = F(z) + \int_0^\infty \exp(-z\theta) Q_\theta \, \mathrm{d}\theta \int_0^\infty \exp(-zt) x_t \, \mathrm{d}t$$

$$= F(z) + F_1(z) \int_0^\infty \exp(-zt) x_t \, \mathrm{d}t,$$

where $F(z)$ is written for $\int_0^\infty \exp(-zt) A_t \, \mathrm{d}t$, and $F_1(z)$ for $\int_0^\infty \exp(-z\theta) Q_\theta \, \mathrm{d}\theta$. Clearly $\exp(-zt) \log x$ tends to zero as t tends to infinity, whilst x never exceeds the initial value $N - y_0$.

Thus:

$$\int_0^\infty \exp(-zt) (z \log x - F_1(z) x) \, \mathrm{d}t = F(z) + \log x_0. \tag{18}$$

It will be seen that this is an equation of the form:

$$\int_0^\infty \phi(x, z) \psi(z, t) \, \mathrm{d}t = \chi(z), \tag{19}$$

where the functions ϕ, ψ and χ are known, and x is a function of t. z may have any value provided that its real part is positive. It follows that the formal solution obtained in the previous paragraph, equation (17), must satisfy this equation (19). If $\phi(x, z)$ had not contained z explicitly equation (19) would be of Fredholm's first type. From this point of view the above equation may be regarded as a generalization of Fredholm's equation of the first type.

Let us now integrate equation (13) with respect to t, between the limits zero and infinity.

We have:

$$-\int_0^\infty \frac{\mathrm{d}\log x}{\mathrm{d}t}\,\mathrm{d}t = \int_0^\infty \int_0^t A_\theta v_{t-\theta}\,\mathrm{d}\theta\,\mathrm{d}t + y_0 \int_0^\infty A_t\,\mathrm{d}t,$$

hence:

$$\log \frac{x_0}{x_\infty} = \int_0^\infty A_\theta\,\mathrm{d}\theta \int_0^\infty v_t\,\mathrm{d}t + y_0 \int_0^\infty A_t\,\mathrm{d}t.$$

If we put A for $\int_0^\infty A_t\,\mathrm{d}t$, and use the relation

$$\int_0^\infty v_t\,\mathrm{d}t = -\int_0^\infty \frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t = x_0 - x_\infty$$

we have:

$$\log \frac{x_0}{x_\infty} = A(x_0 - x_\infty) + Ay_0 = A(N - x_\infty).$$

Let us introduce the value $p = (N - x_\infty)/N$, so that p is the proportion of the population who become infected during the epidemic.

Then $x_\infty = N(1-p)$ and:

$$-\log \frac{1-p}{1-\dfrac{y_0}{N}} = ANp. \tag{20}$$

This equation determines the size of the epidemic in terms of A, N, and y_0, and we shall make use of it later.

If we treat equation (15) in a similar manner, we obtain the relation:

$$\int_0^\infty y_t\,\mathrm{d}t = Np \int_0^\infty B_\theta\,\mathrm{d}\theta.$$

Thus $\int_0^\infty B_\theta\,\mathrm{d}\theta$ is the average case duration.

Finally the observational data are given in terms of x, y and z, though in particular instances the information may be incomplete. The problem may arise of obtaining A_θ and B_θ as functions of θ, and thus of acquiring knowledge regarding ϕ_θ and ψ_θ, the infectivity and removal rates.

In equation (13) v_t and $\mathrm{d}\log x/\mathrm{d}t$ are known functions of t and so the equation is of the type discussed by Fock (1924). We shall apply his method to obtain the solution of this and similar equations.

By equation (13):

$$-\int_0^\infty \exp(-zt)\frac{\mathrm{d}\log x}{\mathrm{d}t}\,\mathrm{d}t = \int_0^\infty \exp(-zt)\int_0^t A_\theta v_{t-\theta}\,\mathrm{d}t\,\mathrm{d}t + y_0\int_0^\infty \exp(-zt)A_t\,\mathrm{d}t$$

$$= \int_0^\infty \exp(-z\theta)A_\theta\,\mathrm{d}\theta\int_0^\infty \exp(-zt)v_t\,\mathrm{d}t + y_0\int_0^\infty \exp(-zt)A_t\,\mathrm{d}t,$$

therefore:

$$\int_0^\infty \exp(-zt)A_t\,\mathrm{d}t = \frac{-\int_0^\infty \exp(-zt)\frac{\mathrm{d}\log x}{\mathrm{d}t}\,\mathrm{d}t}{y_0+\int_0^\infty \exp(-zt)v_t\,\mathrm{d}t}, \tag{21}$$

and we shall denote this last expression by the symbol $F_2(z)$ whence:

$$A_\theta = \frac{1}{2\pi i}\int_{a-i_\infty}^{a+i_\infty}\exp(zt)F_2(z)\,\mathrm{d}z. \tag{21a}$$

By equation (15):

$$\int_0^\infty \exp(-zt)y_t\,\mathrm{d}t = \int_0^\infty \exp(-zt)\int_0^t B_\theta v_{t-\theta}\,\mathrm{d}\theta\,\mathrm{d}t + y_0\int_0^\infty \exp(-zt)B_t\,\mathrm{d}t,$$

whence:

$$\int_0^\infty \exp(-zt)B_t\,\mathrm{d}t = \frac{\int_0^\infty \exp(-zt)y_t\,\mathrm{d}t}{y_0+\int_0^\infty \exp(-zt)v_t\,\mathrm{d}t}, \tag{22}$$

we shall denote this last expression by $F_3(z)$, and so:

$$B_\theta = \frac{1}{2\pi i}\int_{a-i_\infty}^{a+i_\infty}\exp(zt)F_3(z)\,\mathrm{d}t. \tag{22a}$$

Equations (21a) and (22a) give A_θ and B_θ in terms of the observable data. If $F_2(z)$ and $F_3(z)$ can be expressed as rational functions of z, then in place of

Laplace's transformation we can use the simpler solution given in the next section.

3. Special Cases

3.1. The earlier stages of an epidemic in a large population. During the early stages of an epidemic in a large population, the number of unaffected persons may be considered to be constant, since any alteration is small in comparison with the total number. Equation (13) becomes:

$$-\frac{dx}{dt} = v_t = N\left[\int_0^\infty A_\theta v_{t-\theta}\, d\theta + A_t y_0\right],$$

where N is this constant population per unit area.

Using Fock's method:

$$\int_0^\infty \exp(-zt)v_t\, dt = \frac{Ny_0\int_0^\infty \exp(-zt)A_t\, dt}{1-N\int_0^\infty \exp(-zt)A_t\, dt}, \tag{23}$$

and we shall denote this by $F_4(z)$.

Thus:

$$v_t = \frac{1}{2\pi i}\int_{a-i_\infty}^{a+i_\infty} \exp(zt)F_4(z)\, dz. \tag{23a}$$

Making use of equation (15) we have similarly:

$$\begin{aligned}
\int_0^\infty \exp(-zt)y_t\, dt &= \int_0^\infty \exp(-zt)\int_0^t B_\theta v_{t-\theta}\, d\theta\, dt + y_0\int_0^\infty \exp(-zt)B_t\, dt \\
&= \int_0^\infty \exp(-zt)v_t\, dt \int_0^\infty \exp(-z\theta)B_\theta\, d\theta \\
&\quad + y_0\int_0^\infty \exp(-zt)B_t\, dt \\
&= \frac{Ny_0\int_0^\infty \exp(-zt)A_t\, dt\int_0^\infty \exp(-zt)B_t\, dt}{1-N\int_0^\infty \exp(-zt)A_t\, dt} \\
&\quad + y_0\int_0^\infty \exp(-zt)B_t\, dt
\end{aligned}$$

$$= \frac{y_0 \int_0^\infty \exp(-zt) B_t \, dt}{1 - N \int_0^\infty \exp(-zt) A_t \, dt}, \tag{24}$$

which we shall call $F_5(z)$.

Thus:

$$y_t = \frac{1}{2\pi i} \int_{a-i_\infty}^{a+i_\infty} \exp(zt) F_5(z) \, dz. \tag{24a}$$

Further we may find the integral equation for y_t as follows:

$$\begin{aligned}
y_t &= \int_0^t B_{t-\theta} v_\theta \, d\theta + B_t y_0 \\
&= N \int_0^t B_{t-\theta} \left(\int_0^\theta A_{\theta - z} v_z \, dz + A_\theta y_0 \right) d\theta + B_t y_0 \\
&= N \int_0^t B_{t-\theta} \int_0^\theta A_{\theta - z} v_z \, dz \, d\theta + N y_0 \int_0^t B_{t-\theta} A_\theta \, d\theta + B_t y_0 \\
&= N \int_0^t A_{t-\theta} \int_0^\theta B_{\theta - z} v_z \, dz \, d\theta + N y_0 \int_0^t A_{t-\theta} B_\theta \, d\theta + B_t y_0 \\
&= N \int_0^t A_{t-\theta} (y_\theta - B_\theta y_0 + B_\theta y_0) \, d\theta + B_t y_0 \\
&= N \int_0^t A_{t-\theta} y_\theta \, d\theta + B_t y_0 .
\end{aligned} \tag{25}$$

It is easy to show that by solving this directly we obtain the solution (24).

In a previous communication, McKendrick (1925/26), these solutions were given in a somewhat different form. The equation for $v_{t,0}$ was given as:

$$v_{t,0} = \int_0^t A_\theta v_{i-\theta,0} \, d\theta,$$

and the solution obtained was:

$$v_{t,0} = \frac{1}{2\pi i} \int_{a-i_\infty}^{a+i_\infty} \exp(zt) \frac{N_0}{1 - \int_0^\infty \exp(-z\theta) A_\theta \, d\theta} \, dz.$$

It was remarked that $v_{t,0}$ had a singularity at the point $t = 0$. In the present

discussion we regard the original infections as occurring at the very beginning of the epidemic but in such a way as to be independent of the equations which define the epidemic proper. Thus $v_{t,0} = v_t$ except in the short interval of time 0 to ε, and during this interval the integral equation does not hold, but instead $\int_0^\varepsilon v_{t,0}\,dt$ is equal to y_0.

Thus:

$$\begin{aligned} v_{t,0} &= v_{t,0} - v_{\varepsilon,0} + v_{\varepsilon,0} \\ &= \int_\varepsilon^t A_{t-\theta} v_{\theta,0}\,d\theta + \int_0^\varepsilon A_{t-\theta} v_{\theta,0}\,d\theta \\ &= \int_0^t A_{t-\theta} v_\theta\,d\theta + A_{t-\varepsilon'} \int_0^\varepsilon v_{\theta,0}\,d\theta \quad \text{where} \quad 0 < \varepsilon' < \varepsilon \\ &= \int_0^t A_{t-\theta} v_\theta\,d\theta + A_t y_0. \end{aligned}$$

Thus the integral equation previously given for $v_{t,0}$ implies the equation now given for v_t. The solution previously given may be written in the form:

$$v_{t,0} = \frac{1}{2\pi i} \int_{a-i_\infty}^{a+i_\infty} \exp(zt) F(z)\,dx,$$

where:

$$F(z) = \frac{y_0}{1 - \int_0^\infty \exp(-z\theta) A_\theta\,d\theta},$$

let us denote this by $y_0/(1-A)$. In the new form:

$$F_4(z) = -y_0 + \frac{y_0}{1-A} = \frac{A y_0}{1-A},$$

which is the same as in equation (23) when one notes that in the former discussion the function A was taken as including N. Now if v_t has no singularities, the Laplacian solution of $F_4(z)$ is a function with no singularities and so the Laplacian of y_0 corresponds to the singularity. It is easy to see that the Laplacian solution $(1/2\pi i)\int_{a-i_\infty}^{a+i_\infty} \exp(zt)\,(-y_0)\,dz$ corresponds to a function $\phi(t)$ such that $\int_0^\infty \exp(-zt)\phi(t)\,dt = -y_0$. Now if $\phi(t)$ is zero from ε to ∞, and becomes infinite at the origin in such a way that $\int_0^\varepsilon \phi(t)\,dt$ tends to y_0 as ε tends to zero, then it is clear that the above equation will be true. And so the expression $\int_{a-i_\infty}^{a+i_\infty} \exp(zt)\,(-y_0)\,dz$ may be taken as representing a function with

exactly the same properties as $v_t - v_{t,0}$. That is to say it is zero from ε to ∞ and $\int_0^{\varepsilon} (v_t - v_{t,0})\,dt = -y_0$, when ε becomes very small.

These values of v_t and y_t constitute the general solution of the problem in the case where N is considered as remaining constant, if A_θ and B_θ, or ϕ_θ and ψ_θ are given.

We can as before readily obtain the values A_θ and B_θ from observed values of v_t and y_t, and we find:

$$A_\theta = \frac{1}{2\pi i}\int_{a-i_\infty}^{a+i_\infty} \exp(zt)\,\frac{\int_0^\infty \exp(-zt)v_t\,dt}{Ny_0 + N\int_0^\infty \exp(-zt)v_t\,dt}\,dz, \tag{26}$$

and:

$$B_\theta = \frac{1}{2\pi i}\int_{a-i_\infty}^{a+i_\infty} \exp(zt)\,\frac{\int_0^\infty \exp(-zt)y_t\,dt}{y_0 + \int_0^\infty \exp(-zt)v_t\,dt}\,dz. \tag{27}$$

For the arithmetical solution of the integral equations the reader is referred to Whittaker (1918).

It will be observed that solutions (21)–(24), (26) and (27) depend upon an equation of the type $\int_0^\infty \exp(-zt)\phi(t)\,dt = F(z)$ whose solution can be expressed by the use of Laplace's transformation.

If $F(z)$ can be expressed as a rational function of the form $\psi_n(z)/\psi_m(z)$ where ψ_n and ψ_m are polynomials of degree n and m respectively, and n is less than m, then it is always possible to express $F(z)$ in the form $\sum\sum\{A_{r,s}/(z-\alpha_r)^s\}$ where r and s vary from unity to a and b respectively, and a and b have finite values.

But:

$$\int_0^\infty \exp(-zt)\exp(\alpha t)t^c\,dt = \frac{c!}{(z-\alpha)^{c+1}},$$

hence a solution of:

$$\int_0^\infty \exp(-zt)\phi(t)\,dt = \sum\sum \frac{A_{r,s}}{(z-\alpha_r)^s},$$

is given by (Fock, 1924):

$$\phi(t) = \sum\sum \frac{A_{r,s}}{(s-1)!}\,t^{s-1}\exp(\alpha_r t). \tag{28}$$

3.2. Constant rates. Much insight can be obtained as to the process by which epidemics in limited populations run their peculiar courses, and end in final extinction, from the consideration of the special case in which ϕ and ψ are constants κ and l respectively.

In this case the equations are:

$$\left.\begin{aligned} \frac{dx}{dt} &= -\kappa xy, \\ \frac{dy}{dt} &= \kappa xy - ly, \\ \frac{dz}{dt} &= ly, \end{aligned}\right\} \tag{29}$$

and as before $x+y+z=N$.

Thus:

$$\frac{dz}{dt} = l(N-x-z),$$

and $dx/dz = -(\kappa/l)x$, whence $\log(x_0/x) = (\kappa/l)z$, since we assume that z_0 is zero.

Thus:

$$\frac{dz}{dt} = l(N - x_0 \exp(-(\kappa/l)z) - z).$$

Since it is impossible from this equation to obtain z as an explicit function of t, we may expand the exponential term in powers of $(\kappa/l)z$, and we shall assume that $(\kappa/l)z$ is small compared with unity.

Thus:

$$\frac{dz}{dt} = l\left\{N - x_0 + \left(\frac{\kappa}{l}x_0 - 1\right)z - \frac{x_0\kappa^2 z^2}{2l^2}\right\}.$$

But $N - x_0 = y_0$, where y_0 is small. It is for this reason that we have to take into consideration the third term in z^2, as although $(\kappa/l)z$ is small compared with unity, its square may not be small as compared with $((\kappa/l)x_0 - 1)z$.

The solution of this equation is:

$$z = \frac{l^2}{\kappa^2 x_0}\left\{\frac{\kappa}{l}x_0 - 1 + \sqrt{-q}\tanh\left(\frac{\sqrt{-q}}{2}lt - \phi\right)\right\}, \tag{30}$$

where:

$$\phi = \tanh^{-1} \frac{\frac{\kappa}{l} x_0 - 1}{\sqrt{-q}},$$

and:

$$\sqrt{-q} = \left\{ \left(\frac{\kappa}{l} x_0 - 1 \right)^2 + 2 x_0 y_0 \frac{\kappa^2}{l^2} \right\}^{1/2}.$$

Also for the rate at which cases are removed by death or recovery which is the form in which many statistics are given:

$$\frac{dz}{dt} = \frac{l^3}{2 x_0 \kappa^3} \sqrt{-q} \operatorname{sech}^2 \left(\frac{\sqrt{-q}}{2} lt - \phi \right). \tag{31}$$

An example is given in Fig. 1.

Further at the end of the epidemic:

$$z = \frac{2l}{\kappa x_0} \left(x_0 - \frac{l}{\kappa} \right), \tag{32}$$

where y_0 has been neglected. This is obviously no limitation as y_0, the initial number of infected cases is usually small as compared with x_0. It is clear that when x_0, which is identical with N if y_0 be neglected, is equal to l/κ, no epidemic can take place. If, however, N slightly exceeds this value then a small epidemic will occur, and if we write $N = (l/\kappa) + n$, its magnitude will be:

$$2 \frac{l}{\kappa} \frac{n}{N} \quad \text{or} \quad 2n - \frac{2n^2}{N}.$$

In this sense the population density $N_0 = l/\kappa$ may be considered as the threshold density of the population for an epidemic with these characteristics. No epidemic can occur unless the population density exceeds this value, and if it does exceed the threshold value then the size of the epidemic will be, to a first approximation, equal to $2n$, that is to twice the excess (if n is small as compared with N). And so at the end of the epidemic the population density will be just as far below the threshold density, as initially it was above it.

At first sight it appears peculiar that in such a homogeneous population the epidemic should at first increase and then diminish. The reason for this behaviour is readily appreciated when attention is focused on the conditions obtaining when the epidemic is at its maximum. By equation (29) this occurs when $dy/dt = 0$, that is when $x = l/\kappa$, or when the unaffected population has been reduced to its threshold value. Once the population is below this value, any particular infected individual has more chance of being removed by

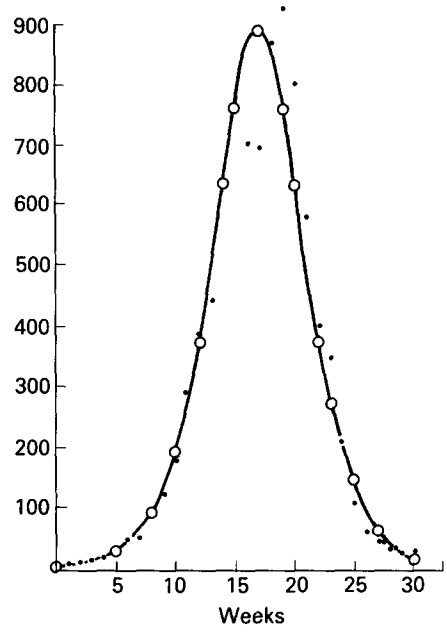

Figure 1. Deaths from plague in the island of Bombay over the period 17 December 1905 to 21 July 1906. The ordinate represents the number of deaths per week, and the abscissa denotes the time in weeks. As at least 80–90% of the cases reported terminate fatally, the ordinate may be taken as approximately representing dz/dt as a function of t. The calculated curve is drawn from the formula:

$$y = \frac{dz}{dt} = 890 \operatorname{sech}^2(0.2t - 3.4).$$

We are, in fact, assuming that plague in man is a reflection of plague in rats, and that with respect to the rat: (1) the uninfected population was uniformly susceptible; (2) that all susceptible rats in the island had an equal chance of being infected; (3) that the infectivity, recovery, and death rates were of constant value throughout the course of sickness of each rat; (4) that all cases ended fatally or became immune; (5) that the flea population was so large that the condition approximated to one of contact infection. None of these assumptions are strictly fulfilled and consequently the numerical equation can only be a very rough approximation. A close fit is not to be expected, and deductions as to the actual values of the various constants should not be drawn. It may be said, however, that the calculated curve, which implies that the rates did not vary during the period of epidemic, conforms roughly to the observed figures.

recovery or by death than of becoming a source of further infection, and so the epidemic commences to decrease. In fact, as remarked above, in small epidemics the curve for y is symmetrical about the maximum. This symmetry exists for y as a function of t, and consequently also for dz/dt, that is to say the curve of removal by recovery or by death. On the other hand no such symmetry is obtained in the curve of case incidence, that is of $-dx/dt = \kappa xy$. This is clear since y is symmetrical and $x = \exp((-l^2/\kappa)\int y\, dt)$.

3.3. Magnitude of small epidemics in general case. We have seen that in the case last discussed, that is where the population is limited, and the characteristic rates are constants, a threshold value exists, such that no epidemic can arise if the density is below this value, whereas if the density be above it, the size of the epidemic is equal to twice the excess, provided that the excess be a small fraction of the threshold density. It is of importance to enquire how far a similar result is true in the general case where the characteristic rates vary during the course of the disease.

We found that:

$$-\log\frac{1-p}{1-\frac{y_0}{N}} = ApN, \tag{20}$$

where p is the proportion of the population infected during the epidemic, and

$$A = \int_0^\infty A_\theta \, d\theta = \int_0^\infty \phi_\theta \exp\left(-\int_\alpha^\theta \psi\alpha \, d\alpha\right) d\theta.$$

We shall assume that y_0/N is small as compared with unity, and can be neglected.

It is clear that when p is greater than zero, $-\log(1-p) > p$, hence $ApN > p$ and consequently $AN > 1$.

That is to say for an epidemic to occur (that is for p to be greater than zero), N must be greater than $1/A$. Writing $N_0 = 1/A$ and $N = N_0 + n$ we have:

$$p + \frac{p^2}{2} + \frac{p^3}{3} + \cdots = ApN$$

$$= p\left(1 + \frac{n}{N_0}\right),$$

hence:

$$\frac{p}{2} + \frac{p^2}{3} + \cdots = \frac{n}{N_0},$$

or neglecting powers of p higher than the first:

$$pN = 2n\frac{N}{N_0} = 2n\left(1 + \frac{n}{N_0}\right) = 2n, \tag{33}$$

approximately, as n/N_0 may be neglected as compared with unity.

A difficulty occurs due to the fact that y_0 can have no value less than unity,

and so y_0/N cannot be made indefinitely small. It appears, in fact, that under certain conditions quite a number of cases might occur at the threshold value, but these would be sporadic cases and would not constitute an epidemic in the true sense. The difficulty may be got over if we allow the unit of area to increase. If we increase it κ times then N_0 increases to κN_0 and A becomes A/κ, so that AN_0 does not change. On the other hand y_0/N_0 becomes $y_0/\kappa N_0$, and although y_0 can never be less than unity, κ can be made indefinitely large, and so $y_0/\kappa N_0$ may ultimately be neglected as compared with unity.

It thus appears that precisely the same result is arrived at in this case, as in the simpler case in which the rates were constants. There exists a threshold population whose density is equal to $1/A$, and when an epidemic occurs in a population of slightly higher density, its size is equal approximately to twice the excess.

It will be seen that the more complex expression A now replaces the simpler fraction κ/l. In fact, when the rates are constant:

$$A = \int_0^\theta \kappa \exp\left(-\int_0^\theta l \, d\alpha\right) d\theta = \kappa \int_0^\theta \exp(-l\theta)\, d\theta = \frac{\kappa}{l}.$$

Reverting to equation (20) it is clear that p can never be equal to unity, as long as N is finite, so that an epidemic can never affect all the susceptible members of a limited population. Of course it has to be recognized that when the population has been reduced to small numbers the equations here given do not strictly hold.

It may also be pointed out that the population density $N_0 = 1/A$ is only a threshold density with respect to initial importations of cases which have just been infected. That is to say the cases present at the commencement of the epidemic are assumed to be of the type $v_{0,0}$, and none are of the types $v_{0,1}, v_{0,2} \ldots v_{0,r}$. It is this limitation which renders it impossible in the general case to identify the threshold population with the number who are still unaffected at the instant when the epidemic reaches its maximum, since at that instant many cases will certainly be not just commencing but will be of the type $v_{0,r}$, and so they cannot be treated as equivalent to those which we have assumed to have been originally introduced. Nevertheless there seems little doubt that by analogy with the simpler case in which the rates were constants, the point at which the epidemic reaches its maximum will, in general, correspond approximately with the point at which the remaining unaffected population has been reduced to the threshold value.

Another point of interest arising from equation (20) is in relation to variations in the infectivity rate. It will be seen that the effect of increasing the infectivity from ϕ_θ to $\alpha\phi_\theta$ is to increase A to αA, and consequently the threshold value N_0 is reduced to N_0/α.

Let $\alpha = 1+\beta$, where β is very small, so that β is the fractional increase in the infectivity.

The new threshold is now $N_0/1+\beta = N_0 - \beta N_0$. Consequently the excess being now βN_0, an epidemic of the size $2\beta N_0$ is to be expected. Thus a small increase in the infectivity rate may cause a very marked epidemic in a population which would otherwise be free from epidemic, provided that the population was previously at its threshold value. On the other hand, if the actual density was below the threshold, no epidemic could occur until the infectivity had been increased to such a degree as to make the threshold value less than the actual density.

It is not difficult to extend these results to such diseases as malaria or the plague in which transmission is through an intermediate host. In this case using dashed letters for symbols referring to the intermediate host we have:

and:

$$\left.\begin{aligned} \frac{\mathrm{d}\log x}{\mathrm{d}t}\int_0^t A'_\theta v'_{t-\theta}\,\mathrm{d}\theta + A'_t y'_0 \\ \frac{\mathrm{d}\log x'}{\mathrm{d}t} = \int_0^t A_\theta v_{t-\theta}\,\mathrm{d}\theta + A_t y_0 \end{aligned}\right\}, \tag{34}$$

whence:

and:

$$\left.\begin{aligned} -\log\frac{1-p}{1-\frac{y_0}{N}} = A'p'N' \\ -\log\frac{1-p'}{1-\frac{y'_0}{N'}} = ApN \end{aligned}\right\}, \tag{35}$$

Neglecting y_0/N and y'_0/N' as before we have to a first approximation:

$$p\left(1+\frac{p}{2}\right)p'\left(1+\frac{p'}{2}\right) = AA'pp'NN',$$

thus:

$$\frac{p}{2}+\frac{p'}{2} = AA'NN' - 1. \tag{36}$$

As p and p' are always positive where there is an epidemic, $AA'NN'$ must be greater than 1, or a true epidemic can occur only when $AA'NN'$ is greater than unity. We thus see that there is no threshold in the sense used in the previous paragraph for either man or the intermediate host separately, but that there

exists what may be called a threshold product $1/AA'$, and this must be exceeded by the product NN' in order that an epidemic may occur.

We shall now suppose that the value of $N' = N'_0$, and that $N = 1/(AA'NN'_0) + n$ where n is not very great compared with $1/AA'N'_0$, thus $N = N_0 + n$.

We observe that if the value N had been N_0, the situation would be such that no epidemic could arise. In fact, the product NN' would have been at its threshold value. If, however, N exceeds this value N_0 by an amount n, and if we regard N_0 as remaining fixed, then under this condition N_0 corresponds to a threshold value in the former sense, and we are considering the case in which this threshold value is exceeded by n.

Eliminating p' from the above equations we have to a first approximation:

$$p = \frac{2n}{N_0} \frac{A'N'_0}{1 + A'N'_0}. \tag{37}$$

Three cases may be considered: (1) when N'_0 is very small, $p = 0$, and a true epidemic will not occur; (2) when $N'_0 = 1/A'$, $pN_0 = n$ (the size of the epidemic is here exactly equal to the excess and the result of the epidemic is to reduce the population to its threshold value); (3) when N'_0 is very great, $pN_0 = 2n$, or to double the excess.

In this case the size of the epidemic is the same as in the simple case previously considered. That this should be so is apparent, when we consider that the assumption that N'_0 is very great, is equivalent to the assumption that the intermediate host is so plentiful that we are dealing with a condition which is practically identical with contact infection.

Further reverting to equation (36) and multiplying both sides by $N_0N'_0$ we have:

$$N'_0pN_0 + N_0p'N'_0 = 2N_0N'_0(AA'NN' - 1).$$

We choose:

$$N_0N'_0 = 1/AA' = \pi_0,$$

where π_0 is what we have called above the threshold product. That is to say, when the populations are simultaneously N_0 and N'_0 there will be epidemic. Then:

$$\begin{aligned} N'_0pN_0 + N_0p'N'_0 &= 2(NN' - N_0N'_0) \\ &= 2(\pi - \pi_0), \end{aligned}$$

where π is equal NN', and we suppose that π is greater than π_0. Now let $\bar{N}$ and $\bar{N}'$ be the populations after the epidemic has terminated, and let $\bar{\pi} = \bar{N}\bar{N}'$.

Then:

$$\begin{aligned}\pi-\bar{\pi} &= NN'-(N-\Delta N)(N'-\Delta N') \\ &= N\,\Delta N'+N'\,\Delta N-\Delta N\,\Delta N' \\ &= Np'N'+N'pN-pNp'N' \\ &= NN'(p+p'-pp') \\ &= N_0N_0'(p+p'-pp')+(NN'-N_0N_0')(p+p'-pp').\end{aligned}$$

If the excess of population is small so that $NN'-N_0N_0'$ is small as compared with N_0N_0', we can neglect the second term. Further, pp' can be neglected as compared with p or p', and therefore:

$$\pi-\bar{\pi}=N_0N_0'(p+p')=2(\pi-\pi_0). \tag{38}$$

That is to say, the difference between the values of the product of populations before and after the epidemic is twice the excess of the product before the epidemic over the threshold product. This equation is exactly analogous to equation (33). Somewhat similar results have been previously obtained by one of us (McKendrick, 1912) in an analogous but slightly different problem.

These results account in some measure for the frequency of occurrence of epidemics in populations whose density has been increased by the importation of unaffected individuals. They also emphasize the role played by contagious epidemics in the regulation of population densities. It is quite possible that in many regions of the world the actual density of a population may not be widely different from the threshold density with regard to some dominant contagious disease. Any increase above this threshold value would lead to a state of risk, and of instability. The longer the epidemic is withheld the greater will be the catastrophe, provided that the population continues to increase, and the threshold density remains unchanged. Such a prolonged delay may lead to almost complete extinction of the population. Similar results, though of a somewhat more complicated form, hold for epidemics transmitted through an intermediate host. In this case, in place of the threshold density we have to consider the threshold product.

4. Summary

(1) A mathematical investigation has been made of the progress of an epidemic in a homogeneous population. It has been assumed that complete immunity is conferred by a single attack, and that an individual is not infective at the moment at which he receives infection. With these reservations the problem has been investigated in its most general aspects, and the following conclusions have been arrived at.

(2) In general a threshold density of population is found to exist, which depends upon the infectivity, recovery and death rates peculiar to the epidemic. No epidemic can occur if the population density is below this threshold value.

(3) Small increases of the infectivity rate may lead to large epidemics; also, if the population density slightly exceeds its threshold value the effect of an epidemic will be to reduce the density as far below the threshold value as initially it was above it.

(4) An epidemic, in general, comes to an end, before the susceptible population has been exhausted.

(5) Similar results are indicated for the case in which transmission is through an intermediate host.

LITERATURE

Fock. 1924. *Math. Zeit.* **21**, 161.
McKendrick. 1912. *Paludism* **4**, 54.
McKendrick. 1925/6. *Proc. Edin. Math. Soc.* **44**.
Ross and Hudson. 1915–17. *R. Soc. Proc.* **92A/93A**.
Whittaker. 1918. *R. Soc. Proc.* **94A**, 367.