5 | THE LAW OF MASS-ACTION IN EPIDEMIOLOGY: A HISTORICAL PERSPECTIVE

Hans Heesterbeek

5.1 | INTRODUCTION

The law of mass-action as a paradigm for describing the contact rate of individuals in a population has been in widespread and heavy use in ecology and epidemiology for almost 100 years. The law roughly states that the rate at which individuals of two types, X and Y, meet is proportional to the product of the (spatial) densities of the respective subpopulations: $\propto xy$.

The law originates in the theory and practice of chemical reaction kinetics. The analogy between individuals meeting when moving around in space and molecules meeting when moving around in a gas or solution is intuitively pleasing. Moreover, the simplicity of the interaction term widens the possibilities to analytically study the behaviour of the systems of differential equations incorporating this description of the contact process. Both of these factors have undoubtedly been major determinants in creating the success of mass action as a modelling concept.

As with many paradigms, its relation to the process it supposedly describes is strained at best. In contrast to the situation in chemical reaction kinetics, there is little evidence that in ecology any form of contact among individuals abides closely to this law. If ecologists scrutinize what they are assuming about the behaviour of individuals and the popula-

tion they constitute, then it is clear that no contact process follows these rules. It is, however, a forceful paradigm if it, despite all its flaws, has had such an effect on the development of ecological and epidemiological theory over a broad range of applications. It is surely one of the most powerful and most useful old metaphors of ecology. The best of these metaphors are easy to grasp, and through their use, profound and robust biological insights can be gained. Picasso once remarked: "Art is the lie that helps us to discover the truth." Ecologists can replace "art" with "modelling" and see the value of this statement when they realize that the law of mass action is the most basic of lies about the contact process in a population.

In this chapter I concentrate mainly on epidemiological mass action. These developments precede the introduction to ecology but not necessarily because one influenced the other. I will show how mass-action was introduced into the study of interactions between susceptible and infected individuals, who was responsible for this, and which of the various "discoveries" of this description of interaction was crucial for the success of the metaphor. I will show that, contrary to what most researchers in epidemic theory today have been led to believe—by chains of papers copying one another's references—the celebrated paper by Hamer from 1906 was not instrumental in the success of mass-action. Mass-action in epidemiology did not originate with Hamer but with Ronald Ross and Anderson McKendrick, who both arrived at the concept simultaneously but by different routes. Only one of these consisted of translating the original chemical definition into interaction among individuals.

Because of its simplicity, and because it made analytical results possible, mass-action has, from the start, made deeper forays into epidemic phenomena possible. Indeed, it can be stated that the metaphor made it possible for the theory to take off. A few particular strands of the mass-action story—the developments started by Ross and, mainly, McKendrick—set in motion a long chain of researchers writing papers and reacting to one another's ideas and led to a long series of developments and insights. In short, mass-action turned epidemiology into a science.

5.2 | CATO MAXIMILIAN GULDBERG AND PETER WAAGE

It is insightful to first go into the origins in chemical reaction kinetics. The law of mass-action was discovered by two Norwegians, Cato

Maximilian Guldberg (1836–1902) and Peter Waage (1833–1900): two brothers-in-law. A wealth of historical information on the law can be found in the book published in 1964 by the Norwegian Academy of Sciences to celebrate the centenary of the original paper (Bastiansen 1964, which includes a facsimile of the 1864 paper). Guldberg was a mathematician who also studied physics and chemistry, and he was, among other things, professor of applied mathematics at the University of Christiania (Oslo) from 1869. Waage studied mineralogy and chemistry, after becoming disappointed in medicine, and was professor of chemistry at Christiania from 1866 (for details about their lives see their entries in the *Dictionary of Scientific Biography*, Gillispie 1972, and the chapter by Haraldsen in Bastiansen 1964). The technical details that follow are taken from Lund and Hassel (1964).

Guldberg and Waage's work on chemical affinities led them to formulate a principle concerning the role of the amounts of reactants in chemical equilibrium systems. They first distilled the "law" from experiments (using barium sulphate and potassium carbonate and substituting to barium carbonate and potassium sulphate) and later gave it a mathematically exact formulation. For a chemical equilibrium of substitution reaction $AB + CD = AC + BD$, they found that, for a given temperature, the product of reactant concentrations is proportional to the product of reaction-product concentrations. The proportionality constant gives the force of the formation, the magnitude of which depends on the force of attraction between the reacting substances. They call this the *affinity coefficient*. Similar statements of such a law had been made before, but this was the first time that the law was expressed clearly. They presented their paper, "Studier over Affiniteten," in 1864 to the Norwegian Academy of Sciences; it was printed in 1865 in the *Forhandlingen i Videnskabsselkabet i Christiania*. Apparently (Bastiansen 1964) the presentation provoked no comment or question from the audience. Perhaps not surprisingly, the paper remained almost unknown for a long time. An extended version of their theory published three years later, although written in French, helped matters little because it appeared in a journal of Oslo University, which was not widely read. Only when Wilhelm Ostwald reconfirmed the law experimentally and presented it in a paper in 1877 did their discovery become widely known. Incidentally, Jacobus Henricus van't Hoff independently discovered the law in 1878, something which happened several times in ecology and epidemiology as we will see later in this chapter.

It is interesting that Guldberg and Waage originally formulated their law in a different way. If the concentrations ("active masses") of the reacting molecules are given by p and q and the affinity coeffi-

cient by k, then they concluded that the interaction should be described by:

$$kp^m q^n. \tag{5.1}$$

They had noted, experimentally, that doubling the masses of the different molecules did not always have a similar result in magnitude. Their reasoning was that in the neighbourhood of two molecules taking part in a reaction, there are other molecules exerting their attracting force. The theory works from the principle that the force between the two reactants is dominating in this arena. Later Guldberg and Waage concluded that these secondary forces could be of the same order of magnitude and they ascribe the difference between reactants to these secondary forces (Lund & Hassel 1964). They conclude that it makes more sense to describe the primary affinity by the simpler relation:

$$kpq. \tag{5.2}$$

It also seemed best to use additional terms for the secondary forces when necessary. This has an interesting parallel in the modern epidemic modelling literature, where mass-action has sometimes been modified to allow exponents different from unity for the densities of infected individuals and susceptibles. One problem with the approach is that a mechanistic basis seems to be lacking. Perhaps a detailed study of the chemical literature might generate a basis for the translation to interaction in populations.

5.3 | WILLIAM HEATON HAMER

Hamer (1862–1936) studied medicine in Cambridge and London. He became Medical Officer of Health of the London County Council in 1912, one of the most important positions in British public health, a position he held until his retirement in 1925 (Greenwood 1936).

Hamer had an interest in explaining epidemic curves (i.e., curves of disease prevalence/incidence against time) and was, like several authors before him at the end of the nineteenth century (e.g., see Ransome 1880), focused on periodic behaviour. To understand the developments, discussions, and papers in that period and the first two decades of the twentieth century, we must understand the context in which the study of epidemiology arose. Briefly, the period starts with the definite proof (in 1877 by Robert Koch and Louis Pasteur) that infectious diseases were caused by living organisms. When researchers first started collecting data in earnest (notably British statistician William Farr), they could

clearly document epidemics and started constructing epidemic curves. These curves turned out to be surprisingly regular in shape. Trying to explain mechanistically the shape of these curves gave rise to many scientific studies in that period. For details see Dietz & Heesterbeek (2005), where we also explain in detail that data collection on infectious diseases, and its analysis, started well before the time of Farr: by Graunt in the seventeenth century and, in the eighteenth century, in relation to, for example, smallpox.

Basically researchers can distinguish two competing approaches to explaining the epidemic curve's shape; I called these *Farr's hypothesis* and *Snow's hypothesis*. They arose because of the limited knowledge about the living organisms causing disease and the confusion that early experiments caused because of this limited knowledge. Farr's hypothesis has its roots in the work of Justus von Liebig, who advocated a purely chemical basis for infection, and asserts that epidemics end because the potency of the causal "organism" decreases with every individual it passes through (Farr 1866). John Brownlee was its main mathematical exponent. Snow's hypothesis asserts that epidemics end because the epidemic runs out of "fuel," that is, the decreasing availability of susceptibles causes the outbreak to end (Snow 1853). Ross was the main mathematical exponent of Snow's hypothesis, and ultimately Brownlee gave up his efforts to counter this theory (see Fine 1979 and Dietz & Heesterbeek 2005).

Hamer was also convinced of the validity of Snow's hypothesis. Hamer published his first paper on the epidemiology of infectious diseases in 1896: a review of the most important contributions to the study of the epidemic curve and cyclic behaviour. Until then, the papers addressing periodicity in infectious disease did so systematically—directed by data—and speculatively but not in a theoretical mathematical setting. There was some mechanistic—more descriptive—reasoning, but this reasoning was not taken to the extreme of testing its consequences either quantitatively or qualitatively by analyzing mathematical models based on these mechanisms. In this respect, Hamer's work is different from that of the authors before him; Hamer applied mathematical reasoning in his 1906 periodicity paper but not, as you will see, from a mechanistic description of underlying processes. Hamer did not have a favourable view of the disciplines of mathematics or biometry, to put it mildly, nor do his other papers contain mathematical modelling.

In 1906, Hamer delivered the so-called Milroy Lectures of the Royal College of Physicians, London, which Whitelegge, among others, had given before him. This was an honour bestowed on Greenwood later. Hamer's three lectures were called "Epidemic Diseases in England: The

Evidence of Variability and Persistency of Type" (Hamer 1906). The papers predominantly described disease *type*, by which he meant a combination of virulence, ability to spread, and overall behaviour of the disease's epidemics; the precise definition, however, remains vague. In the third lecture on March 8, 1906, in what is basically a single page (p. 734) of a long paper, the paragraphs of interest appear. The argumentation and theoretical style is in disharmony with the flowing, talkative style of the rest of the papers, which are more like the nineteenth century writing on epidemic theory in England. To strengthen this disharmony, the vital parts of page 734 are printed in a smaller typeface (footnote style), as if it needed to be accentuated that here a rather long side-remark was being made that could be skipped if so desired. Little did he realize that this part of the manuscript would become the basis of his lasting claim to fame.

Hamer starts by stating: "The persistency of type displayed by measles and small-pox is quite remarkable. For that reason they afford specially promising material for study of short-term period waves." The simplest case, he continues, is that of the short-period waves of measles. He lists several observations that might be relevant:

1. "Explosions in towns occur commonly at about biennial intervals when the accumulation of susceptible persons is sufficient and the climatic and other internal conditions offer sufficiently small resistance . . . The mean seasonal wave shows two maxima."
2. "The problem (in the case of measles in a large community) is simplified [because] we are dealing with an obligatory parasite. Hence questions of saprophytic growth, of food outbreaks, &c, do not arise."
3. "Furthermore, one attack confers almost complete protection."
4. "Again, infection spreads from person to person, population being densely aggregated and new susceptible material added in sufficient quantity and with sufficient frequency to favour stable epidemic movement."

The aim of Hamer is to construct an epidemic curve based on several assumptions about transmission and to compare the resulting curve with measles data from London. He is interested in whether such a curve could at least be qualitatively similar to observed patterns. Unfortunately, neither the data nor a curve based on them is given in the paper. Presumably it concerns data from 1880 to 1884 as reported by Whitelegge in 1892. Continuing in the main text, he states: "I have taken the London figures and assumed a case-mortality of 2.5 per cent." He proceeds to explain how he arrived at an archetypal curve for the London measles data, averaging over the curves of an unnamed number of

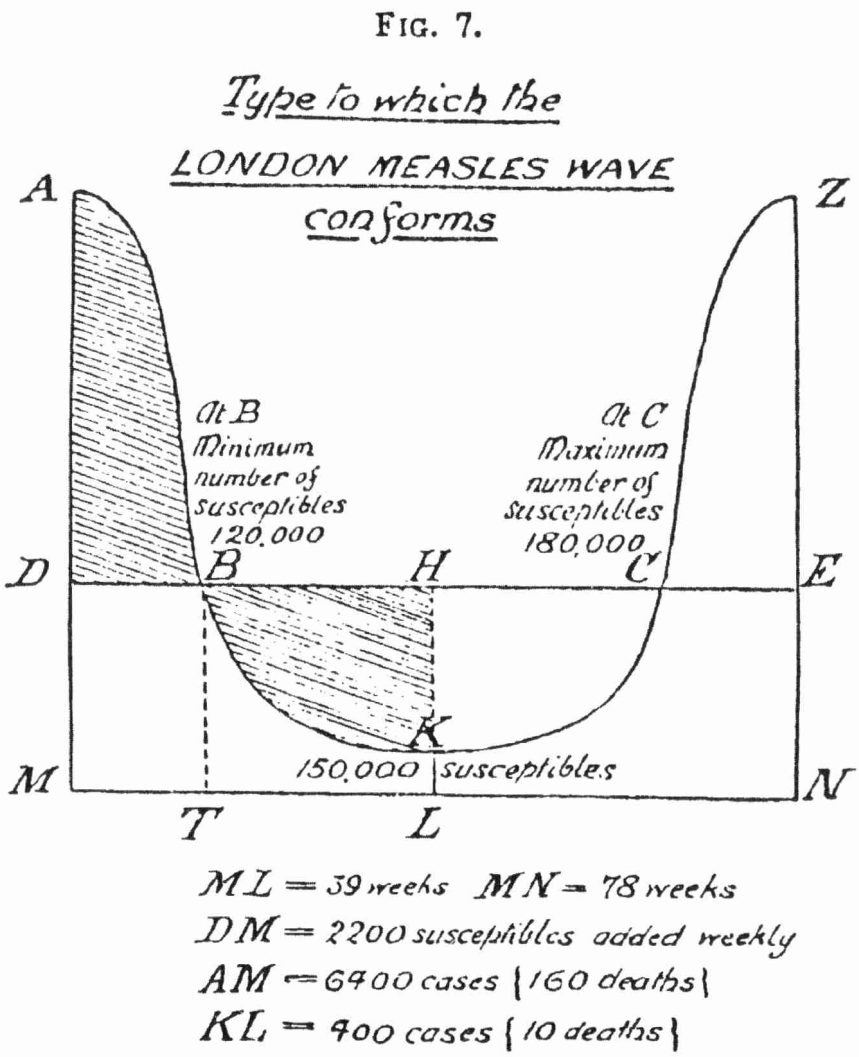

FIGURE 5.1 | Idealization of London measles wave by Hamer; see the main text for an explanation. Figure taken from Hamer 1906, Lecture III, page 734.

epidemic cycles. At this point in Hamer's text the small print starts. The curve is reproduced in Figure 5.1.

I will highlight the main ideas and refer to Figure 5.1 for the explanation of the various special points on the curve. The time-axis *MN* of the typical measles wave in Hamer's view equals 78 weeks. He takes the incubation period of measles to be 2 weeks and considers this the length of the time steps in his curve. He is interested in calculating the number of

new cases arising in the next such interval from the cases in the present interval, i.e., discrete time incidence. In doing this he makes several important remarks. First, concentrate on the left half of Figure 5.1. Hamer notes that at points *A* and *K* the increase (decrease) in the number of cases changes into a decrease (increase):

> It will be apparent, therefore, that at *A* each case may be regarded as infecting one other case; this will also hold good at *K*.
>
> If the virulence of the measles organism and other factors be assumed to be the same at *A* and *K*, it will follow, inasmuch as each case is then capable of infecting one other case, that the number of susceptible persons in the population at those points of time will be identical.

Hamer notes that at points *B* and *C* it must hold that the number of new cases in the week-intervals containing these points equals the total number of new susceptibles added to the population: "Further, area *ADB* = area *BHK*, for the former represents the excess of persons attacked over susceptibles newly added in the time *MT*, and this must be equal to the excess of these newly added over those attacked in the time *TL*." He does a quick estimation using the area of the triangle *ADB*—knowing that at *B* the incidence is about 2200 and at *A* about 6400 with 14 weeks from point *M* to *T*—to arrive at an estimate of 30,000. Upon the assumption that the availability of susceptibles is the only mechanism at play—Snow's hypothesis—he comes to the heart of the matter:

> In the neighbourhood of *B* the cases fall in a fortnight from about 2500 to 2000. Assume the number of susceptibles at *A* to be x, the number at *B* will be $x - 30.000$. If the lessened ability to infect at *B* be solely due to diminution in number of susceptibles we may write
>
> $$\frac{x - 30.000}{x} = \frac{2000}{2500}$$
>
> i.e., $x = 150.000$ approximately.

Examine the crucial equation more closely, which Hamer neglects to do. In words, the relation states the following:

$$\begin{aligned}&[\text{cases in next interval}]/[\text{current cases}]\\&\quad = [\text{current susceptibles}]/[\text{susceptibles in which}\\&\qquad \text{one infects one}].\end{aligned} \tag{5.3}$$

This can be slightly rewritten as follows:

$$[\text{cases in next interval}] \propto [\text{current cases}][\text{current susceptibles}]. \quad (5.4)$$

Here, a proportionality constant is characteristic for the infection and the community and is taken to be the inverse of the steady-state susceptible population size when one infective gives rise to exactly one new case; in Hamer's words, the inverse of "the susceptible population when one infects one." Note that, in modern terms, Hamer shows great understanding here. If the infection is endemic in the community, then each case generates exactly one new case during its infectious period, that is, the reproduction ratio $R = 1$. In a homogeneously mixed population with a constant force of infection, the basic reproduction ratio R_0 can be estimated as the inverse of the endemic susceptibles (e.g., see Diekmann & Heesterbeek 2000). The proportionality constant therefore equals the basic reproduction ratio. Because Hamer takes the infectious period to be of length 1 (i.e., takes the time step equal to the average infectious period), the proportionality constant equals the transmission rate constant of the basic SIR-epidemic model. Even though Hamer did not think in these terms, the words he uses clearly indicate that he should be credited with substantial insight into the matter.

Hamer also indicates full understanding of the problems in combination with the distinction between Snow's and Farr's hypothesis of epidemic decline. He writes, still in small print, the following:

> This examination shows the absurdity of assuming that an epidemic comes to an end because all the susceptibles have been attacked; or, again, of expecting to find explanation of the decline in loss of virulence of the organism or of its infecting power . . . The measles curve just defined sufficiently indicates that an epidemic may come to an end despite the existence of large numbers of susceptible persons in the population, merely on a "mechanical theory of numbers and density," and that the assumption of loss of virulence or infecting power on the part of the organism is quite unnecessary.

Because of its importance in chemistry, the law of mass-action had become a standard part of the education of many students in natural sciences and medicine at the end of the nineteenth century. Two key initiators in ecology and epidemiology, Lotka and McKendrick are known to have had access to such textbooks. Hamer, considering his fine education, could easily have come into contact with the law during chemistry courses. It is certain that he did so before 1906, because in his measles paper the words "mass-action" appear frequently

(see a later section). However, contrary to what is commonly believed, nothing in Hamer's paper suggests that he made the link from chemical mass action to epidemic theory. Indeed, the words mass-action in his paper might have given the impression that he had made the link, but this is a misreading of Hamer's text. There are several arguments to support the view that Hamer failed to "put two and two together". First, he does not refer to mass-action or even chemical kinetics, however vaguely, in or near the small print in which his theory is unfolded, whereas "mass action" occurs frequently later in the same paper in a different context. Second, note that Hamer formulates his relation in vaguer terms than I have given previously and without the verbal interpretation that Soper would give to it 23 years later. Hamer introduced his relation in the direct neighbourhood of two particular points on his curve, *B* and *C*, solely to calculate the minimum and maximum number of susceptibles present during the passing of the epidemic wave. He does not, contrary to Ross, McKendrick, and Soper (see the next sections), introduce his relation based on assumptions that describe how individuals make contact and how frequent that contact is.

A third and important argument follows from the context in which Hamer *does* use the words "mass action." This is a context far removed from the study of measles periodicity. Following his brief theoretical small-print excursion, the remaining two-thirds of Hamer's third lecture is devoted to a purely epidemiological discourse on persistence and variability of type. On page 737 he embarks upon a description of the emerging ideas concerning "immune bodies, in number untold, each specific to its particular toxin." He predicts that the questions related to "bacilli, immune bodies, &c [will receive] a good deal of further attention in the near future." He then describes the "conception of the bacillus as a kind of host with attached 'enzymes.'" These particular "enzymes" (Hamer's quotes) would be capable of attacking the variety of sugars that could be fermented by bacteria. He writes: "We may regard the enzyme . . . as in a condition of equilibrium, influenced equally but in opposed directions by two 'mass actions.'" Later he uses the phrase "mass-action-equilibrium point" several times and he draws analogies between the enzyme example and immune reactions. Of the previous object of his study, however—where the mass-action concept would also be appropriate—no trace is left. Towards the end of the paper Hamer becomes increasingly involved in his argumentation against bacteriological epidemiology, maybe even up to a point where he would be unable to make the connection that Soper in 1929 and others later supplied. For more information on Hamer's life and work I refer to the obituary by Greenwood (1936).

5.4 | RONALD ROSS AND ANDERSON McKENDRICK

Even though Hamer's paper is almost invariably credited with introducing chemical mass action into epidemiology, we have seen that Hamer did not realize he had done so. More remarkably, the credit does not hold in a more fundamental sense: The researchers that, with hindsight, turned out to be the main players in shaping epidemic theory in the early phase of the nineteenth century show no knowledge of Hamer's paper. Ross and McKendrick never refer to Hamer in their work. Although it is true that in those days precise references, in the way we are accustomed to today, were scarce, both Ross and McKendrick are positive exceptions in that they consistently at least give the names and in most cases the journals and page numbers of the literature they use or describe. Both Ross and McKendrick were medical doctors, and it is therefore not unlikely that they had regular access to *The Lancet*, the journal that published Hamer's three papers. Ross and McKendrick performed much of their duties abroad (India, Mauritius, Africa, Sierra Leone), however, and their main interest was in diseases and infections from the tropics. Of McKendrick, for example, it is known that in 1905 and 1906 he was at the Pasteur Institute in Kasauli, India. Even if Ross and McKendrick had seen the papers, it is possible that they did not read the crucial part three because the first two instalments, dealing as they did with measles, are unlikely to have caught their attention. Hamer's paper did not start modern epidemic theory.

If it was not Hamer, who then introduced mass-action into modern epidemic theory? Ross (1857–1932) introduced his "theory of happenings" in the appendix to his famous book *The Prevention of Malaria*, but only in the second edition. Later he referred to the theory as the field of "*a priori* pathometry" (Ross 1916). For a description of the life and work of Ross see, for example, the most recent of several biographies devoted to him (Nye & Gibson 1997). None of the many sources on Ross, however, celebrate in detail his accomplishments as one of the two founding fathers of modern epidemic theory (the other being McKendrick, as explained later). A positive exception is the paper by Paul Fine (Fine 1975). For some details of his later work and its role in the genesis of the basic reproduction ratio, see my article (Heesterbeek 2002). Ross was the first to present a mechanistic theory of epidemics without having a specific infectious agent in mind. He used the term *a priori* to signify that he first makes assumptions about the way infections spread among individuals (through various "happenings"). This contrasts with a widely used approach at that time, for example, in the elaborate work of

Brownlee (see Fine 1979), of trying to gain insight into the phenomenon of epidemics by studying incidence curves of past outbreaks. The theory of Ross from 1916 is based on the 1911 appendix, but there the modelling uses discrete time evolution, whereas in 1916 he switches to ordinary differential equations (and in later papers with Hilda Hudson also integral equations, laying the groundwork for Kermack and Anderson McKendrick). In the 1916 version, Ross writes (p. 208; cf. p. 656 of Ross 1911): "The problem before us is as follows. Suppose that we have a population of living things numbering P individuals, of whom a number Z are affected by something (such as a disease), and the remainder A are not so affected; suppose that a proportion $h.dt$ [sic] of the nonaffected become affected in every element of time dt."

Transmission is thus described as $-hdt.A$ [Ross's notation]. Ross analyses various choices in the resulting systems of equations for P, A, and Z (which also include terms for recovery, birth, death, and migration), and he first devotes ample space to the cases where h is constant. He considers his theory widely applicable because for the constant case he has in mind "such happenings as many kinds of accidents and noninfectious diseases due to causes which operate, so to speak, from outside." Then on page 220 (p. 666 in 1911) he starts the analysis of the case of *dependent happenings* and first considers a *proportional happening*. To the class of dependent happenings, he considers to belong "infectious diseases, membership of societies and sects with propagandas, trade-unions, political parties, etc., due to propagation from within, that is, from individual to individual." For proportional happenings he then writes $h = cZ$. Because he has $Z = P - A$, transmission is then described as $-cAZ = -cA\ (P - A)$. We see that Ross indeed introduces mass-action in this paper, as one of the possibilities of describing transmission, but like Hamer he shows no awareness of this. Ross's mass-action comes as a natural step in the theory he is developing and not as a translation of chemical reaction kinetics.

Incidentally, in the 1917 paper by Ross and Hudson, the follow-up to Ross's 1916 paper, the authors extend the theory by allowing infectious individuals to recover into an immune state and present the model that is nowadays commonly referred to as the *SIR model*. In modern literature, the seminal paper of Kermack and McKendrick from 1927 is usually credited as the first description of the SIR model (even though they only used it as a special case of a more general model in terms of integral equations). Ross and Hudson, however, were also not the first. The credit for the SIR model, that is, the mass-action transmission term and an exponentially distributed infectious period after which the indi-

vidual becomes immune to renewed infection, should go to McKendrick alone in a paper published in 1914. That paper is important for another reason: its main subject is the first clear description of an age-structured model. In a paper from 1926, McKendrick also introduces the first description of the stochastic SIR model (reprinted in Kotz & Johnson 1997 with a commentary by Dietz). It is time to look at McKendrick in more detail.

McKendrick (1876–1943) was, with Ross, by far the most prolific and influential of the pioneers in epidemic theory. For a description of his life and work (and an almost complete list of 50+ publications) see the obituary by W.F. Harvey (1943) or the chapter by Aitchison and Watson (1988) in a book about the influence of Scottish medicine. McKendrick was, like Ross, a medical doctor and a self-taught mathematician. It is without doubt that McKendrick was heavily influenced by the older scientist Ross in his interest in applying mathematical reasoning to medical problems. McKendrick served under Ross during an antimalaria campaign in Sierra Leone and they returned home together by boat in the summer of 1901. Correspondence between the two men exists in the Ross Archives at the London School of Hygiene and Tropical Medicine. From this it is evident that Ross valued highly his own efforts to establish a general mechanistic mathematical theory of epidemics and that he saw in McKendrick the perfect student to carry further what he was starting. In 1911 he wrote in a letter to McKendrick: "We shall end by establishing a new science. But first let you and me unlock the door and then anybody can go in who likes."

During their stay in Sierra Leone and their trip home, Ross must have advised several books to study, because McKendrick thanks him for this in a letter in late 1901. It is not very likely that the book, *Higher Mathematics for Students of Chemistry and Physics, with Special Reference to Practical Work* by Joseph William Mellor, was also on Ross' list (the first edition appeared in 1902). This book was to become McKendrick's "bible" for learning mathematical techniques. It is known that he studied it in great detail while he was stationed in India. The book, which went through many editions in several decades, contains a detailed exposition of the theory of chemical reaction kinetics as it had unfolded by then and included the original chemical law of mass-action. In the introduction, where Mellor lists several uses of mathematics in physical chemistry, he writes: "Wilhemy's law of mass action prepares us for a detailed study of processes of integration" (Mellor 1905). Wilhelmy studied monomolecular reactions. Mellor presents chapters on differential and integral calculus, probability theory, Fourier's theorem, cal-

culus of variations, infinite series, and differential equations. In bold face he remarks in the latter chapter: "A differential equation, freed from constants, is the most general way of expressing a natural law." The author proceeds to treat chemical reactions of the second order in detail and mentions Guldberg and Waage. He writes: "When the system contains $a - x$ gram molecules of acetic acid it must also contain $b - x$ gram molecules of alcohol. Hence $dx/dt = k(a - x)(b - x)$."

It is therefore not surprising that McKendrick put "two and two together" and considered it natural to use the chemical description as a metaphor for interaction between susceptible and infectious individuals. In 1910 McKendrick worked on a mathematical theory of "serum dynamics", by which he meant the "multitude of phenomena of widely varying character" such as agglutination, bacteriolysis, and haemolysis. He then wrote (McKendrick 1911): "It is the object of this paper to show that the above reactions are subject to the law of mass action." The content of the paper is not of importance to epidemic theory directly, but it shows how McKendrick thought about modelling interaction phenomena. That he considered the law of mass-action of universal use becomes clear in a paper read before a conference on malaria. Where Ross initially used discrete-time systems for his theory of happenings, McKendrick, under the influence of Mellor, used differential equations from the start. Ross turned to differential equations in a paper on malaria in *Nature* in 1911 (this paper contains both the discrete time and the continuous time formulation), and McKendrick remarks in the conference proceedings in *Paludism* in 1912 that "he was glad to see . . . that his [Ross's] final results, as there expressed in the form of differential equations, were the same as his own." He does not follow in Ross's footsteps and ignores the approach through "happenings". He writes:

> I propose to develop the subject on my own lines as the argument used is easier to understand and I think more convincing. The mathematical classification of epidemic diseases is a very different classification to that arrived at from the medical standpoint. I will not deal with this in detail here, but I will in the first instance consider a type of epidemic which is spread by simple contact from human being to human being and in which there is no recovery rate. An example of this would be an epidemic of itch in a fixed population of, say, guinea pigs.
>
> The rate at which this epidemic will spread depends obviously in the number of infected animals, and also on the number of animals which remain to be infected—in other words the occurrence of a new infection depends on a *collision* [my italics, H] between an infected and an uninfected animal. If y be the number infected, and a be the total popula-

> tion, then $a - y$ = the number uninfected. If we denote rate of increase of infected by the symbol dy/dt then we have *at once* [my italics] the equation
>
> $$\frac{dy}{dt} = ky(a - y).$$
>
> Where k is a factor which measures the chance of infection, it includes degree of dispersion of individuals, degree of intercourse, and the chance of transmission, etc.

The use of the word *collision* clearly shows that McKendrick considers this interaction to be similar to chemical mass-action (later in the paper he refers to it as "the argument of collisions"), and the words *at once* indicate that he considers this analogy to be a natural one. This seems to be the first clear statement of chemical mass action as applied to the interaction among individuals in the epidemiological-ecological context. Given that, in addition, it was McKendrick's work, rather than Hamer's, that influenced the evolution of epidemic theory as we know it today, I feel future credit for introducing mass action into epidemiology and ecology should go to McKendrick.

In addition to Ross and Hudson's and McKendrick's work, there was another important paper in which mass-action was used in an epidemic, or rather endemic, model. In 1921 Martini published a paper in which he used the mass-action formulation for an infection leading to immunity, where births and deaths of hosts are taken into account. The resulting equations were analyzed by Lotka (1923) and treated in his influential book (Lotka 1925). Lotka's initial exposure to epidenic theory came from reading the work of Ross.

5.5 | HERBERT EDWARD SOPER

There was at least one scientist with a mathematical education who did familiarize himself with Hamer's paper. Soper (1865–1930) was originally an engineer who became interested in statistics while studying under Pearson in London. In 1922 he published a book, *Frequency Arrays*, on generating functions (for details about this book and a few details about the life of Soper, see the obituary by Greenwood in 1931), and shortly after he started work as a mathematician in Brownlee's department at the National Institute of Medical Research. The aim was that he should bring some structure to biometry and "give mathematical form to

Brownlee's own doctrines" (Greenwood 1931). Brownlee had little interest in others' work; as a consequence, Soper did not receive much encouragement to develop his own views on epidemic theory. Given the content of his 1929 paper, it is clear that there would have been ample material for discussion because Soper's view of epidemics was, like Hamer's, based on Snow's hypothesis rather than Farr's hypothesis, of which Brownlee was the main advocate. It was only after the death of Brownlee in 1927, when Soper nominally passed to Greenwood's department in the same institute, that Soper received the freedom and credit he deserved. In Greenwood's department he wrote his important paper on measles periodicity, to be published in 1929. Unfortunately, there was little time to further develop his theory, because Soper died shortly after that publication, on September 10, 1930.

Although Soper's paper appeared more than 10 years after the Ross-Hudson papers, it should be treated independently because Soper does not show any knowledge of the earlier abstract work. Ross is not mentioned, either by Soper or by any of the discussants to the paper (described later in this chapter). In addition, Soper does not appear to have had any knowledge of the first Kermack-McKendrick paper, published 2 years earlier, or of the elaborate preceding work by McKendrick. It might, like in the case of Ross and McKendrick remaining oblivious of Hamer's work, be simply a case of working in two fields perceived as different: tropical infections that mainly affected the colonies and childhood infections that still affected the United Kingdom. Had Soper read any of these papers he could have formulated a large part of his theory less awkwardly.

Soper's aim is to "adopt the simplest mathematical postulate that would describe in a first measure the generally accepted mechanism of epidemic measles, if the accumulation of susceptibles were really the prime factor" and then to "compare the deduced results with the observed facts" and, if necessary, "to modify the primary hypothesis". In a nutshell, Soper describes a philosophy for research in epidemic theory in that he combines inferences from dynamic modelling with data to scrutinize the assumptions underlying the original, simple model, concluding that they need adjustment and then improving predictions by trying a slightly less simple version of the model. Soper was probably the first with this approach in the English literature; Ross and Hudson do not apply their theory to any data and Kermack and McKendrick merely use their single example as an illustration; they do connect to data in their later papers. Effectively, the honour for the first to devise an abstract mathematical model, to confront it with data and then to modify the model, should go to Piotr Dmitirievich En'ko (1844–1916), but unfortu-

nately he published in Russian (in 1889). See Dietz (1988) for a review of his work and En'ko (1989) for a translation by Dietz of (the main part of) the paper from 1889.

Soper is humble in that he states he is "merely following up the trail blazed by Sir William Hamer . . . only in detail departing from his methods." The hypotheses on which Soper starts his investigation are clearly stated. He assumes that in a certain population there is "a perpetual flow of susceptibles possessing three characteristics, *viz.* (1) an equal susceptibility to a disease prevalent in the community, (2) an equal capacity to transmit the disease according to a law, when infected, and (3) the property of passing out of observation when the transmitting period is over." The chosen "law" is the key element that determines the course of the epidemic. Soper clearly wants to separate the law in two distinct processes. The first concerns the infectious period and infectivity, and the second concerns the opportunities for transmission. For the infectivity part, he presents a complicated-looking argument, which comes close to Kermack and McKendrick's idea of a general infectivity function. In essence he states that the "power of transmitting infection is some function of the lapse of time from a definite infection instant" and he illustrates this by drawing the typical unimodal function of measles infectivity. Soper initially chooses the extreme view that "all infecting power is concentrated" at the "definite end" of the incubation period, constituting "an instantaneous power of reinfecting". Later in the paper he assumes a geometric distribution for the infectivity function, expressing the long-popular alternative of a constant infectiousness for an exponentially distributed period.

Having concretized the infectivity, Soper turns to the contact opportunities: "The instant or point infection law being accepted, it is next assumed that a process analogous to 'mass action' governs the operations of transmission and that, other things equal, the number of cases infected by one case is proportional to the number of susceptibles in the community at the instant." The time unit chosen is 2 weeks. If zdt are the cases arising per unit of time, Soper expresses Hamer's "formula" as:

$$\frac{z_{1/2}}{z_{-1/2}} = \frac{x}{m}, \tag{5.5}$$

where $z_{1/2}$ is the number of cases in the unit interval succeeding the present instant. Unlike Hamer, Soper tries to give more meaning to the parameter m. Like Hamer, he refers to it as the "steady-state" number of susceptibles in which "one infects one". However, he then states: "Since

the synthetic epidemics to be made do not depend on absolute sizes of communities, we relate m and a by the quotient s defined by $m = sa$ and think of a community as characterized by a time element s . . . The space of time is a measure of the 'seclusion' of the community, a large arguing few interminglings of the sort that conduce the infection." Here we can see that both Hamer and Soper had in mind that the constant should be a measure of the contact opportunities between susceptibles and infectives in a community.

By assuming that $a = 1000$ per 2-week interval, and taking $s = 20$, 30, 40, and 50, Soper generates a "series of epidemic curves showing all the features found by [Hamer] among them the asymmetry. I do not find any damping, and the curves appear to repeat themselves precisely." He proceeds to calculate the period of the oscillations and finds that it is $2\pi\sqrt{s\tau}$ "under the laws taken", in which τ is the length of the incubation period. For the London data that Hamer used, Soper is reasonably satisfied with the results. However, "the Glasgow curve of measles cases 1888–1927 . . . [does] not show anywhere the simple form of single repeated wave, and Glasgow appears to be rather different from London . . . in the course taken by its measles epidemics."

The only way in which he could create something similar to Hamer's composite curve was to average the epidemic over six consecutive 2-year periods, but the resulting curve still showed "a large winter peak and a small summer forepeak". Although retaining his assumption that the infectious period is a point event, Soper investigated two additions to his model to see whether the Glasgow curves could be recreated. First, he introduced a factor k_θ "representing the influence of the season θ . . . such as might be brought about by school break-up and reassembling." In addition, he took the inflow A (equal to a, which is part of the constant m) of new susceptibles in each time interval (incubation period) into account. He obtained:

$$\frac{z_{1/2}}{z_{-1/2}} = k_\theta\left(\frac{x+A}{m}\right), \tag{5.6}$$

where the incubation period was chosen as 1 month for convenience (Soper raised the left-hand side of the equation to the power i, which is the incubation period in months; he referred to the expression with $i = 1/2$ as the "infectivity"). He was not happy with the "makeshift" procedure he devised for estimating A, i, and k_θ from data and was—after detailed calculations of periods and resulting curves—not happy with the final predictive power of his equation: "The law of propagation of the

disease of measles . . . is therefore not quite so simple that we can get good forecasts merely by premising a uniform inflow of susceptibles, who will all take the disease, and an infectivity depending on the accumulation of such susceptibles and on a season factor." Several factors are responsible for the poor quality of the predictions, according to Soper. Among them is, for example, the assumption of "an even inflow of susceptible persons who will become registered cases" and that "susceptibility cannot be thought of as a unitary character, but certainly varies with age". The most important point, however, "is, perhaps, a false analogy between infection in disease and the mechanism understood under the name of chemical mass action." Soper continues:

> Apart from the great differences in the statistical numbers dealt with in the two fields, in a liquid the intimate uniformity of the mixture and the conditions of intermingling and collision are likely to be more law-abiding than are similar traits in a community of persons. After all, the contacts are comparatively few and are subject to volition as well as to chance and, in addition, a single community may be quite differently constituted in respect to its seclusion or minglings in one part or another.

There may be different degrees of mingling in different sections of the community or depending on the season. Even if "something approaching the mass-action law of infection" is appropriate in each of these sections and seasons, "it is still questionable how far the different curves with their different phases and, perhaps, different periods will unite into a single curve having the same characteristics." He added that it is perhaps "the imperfect mixture and the imperfect tuning of the parts [that is] responsible for the apparent discord in the whole." It is perhaps ironic that the man who would finally provide a mathematical foundation of Hamer's ideas and would link these ideas to chemical mass-action came to the conclusion that the mass-action assumption was not realistic for epidemiology. It is a pity that Soper was not trained more as a modeller.

As is usual with papers "read" to the Royal Statistical Society, Soper's paper is followed by an elaborate commentary (12 pages). The paper is met with great acclaim from all commentators. Some of the criticism could have been valuable to Soper in steering his research. However, because Soper died soon after publication, he did not have the opportunity to follow any of the suggestions made. Hamer, who was in audience, called the analysis "extraordinarily interesting" and likened the transformation of his original idea to Soper's theory to the transforma-

tion of Eliza Doolittle into "one of the most peerless Galateas that ever stepped off a pedestal" in Bernard Shaw's *Pygmalion*. Neither Ross nor McKendrick, who never published in statistical journals, are discussants of Soper's paper and, more remarkably, are not mentioned by any of the discussants, even though both published papers in which similar problems were studied in a similar way. Greenwood knew of Ross's work and could have stimulated Soper to try and merge a statistical and a modelling approach. Greenwood (1916), in basically a review of the work by Ross and Brownlee, wrote (describing only the model of Ross with constant happenings, not the mass-action-type model):

> The advantage of Sir Ronald Ross's method, apart from its simplicity and elegance—advantages which are, however, no mean matters—lies in its generality, so that it may be possible to include the case hypothesized by Brownlee as a particular example . . . As restrictions are relaxed, the analysis will inevitably become more intricate, and, having devised an *a priori* law, one must devise . . . a way of applying the law to statistical data.
>
> It is high time that epidemiology was extricated from its present humiliating position as the plaything of bacteriologists and public health officials . . . The work of Sir Ronald Ross, of Dr. Brownlee, and of a few others should at least elevate epidemiology to the rank of a distinct science.

At this time, Greenwood showed no knowledge of the equally relevant work by McKendrick, nor did he do so in later publications (such as in Greenwood 1932). Ross did not publish on mathematical epidemiology in the period (he died in 1932), so we do not know whether Ross read Soper's paper. McKendrick shows knowledge of the paper in later work. He referred to it in an article in 1940 but only as the source of a graph of the Glasgow measles periodic curve. (It is something of an honour, however, because Soper is the only reference of 20 that is not a work by McKendrick!) Although the situation has improved, it is still the case that statistical and mathematical modellers of epidemic phenomena form two groups with far too little overlap and far too little understanding and knowledge of each other's results, theories, and problems. Clearly, the languages of the two groups and the problems they address have diverged, but strangely enough even in the early twentieth century, when the material was still similar, bifurcation occurs. One explanation might be that the approach by Ross and McKendrick was *a priori*, starting with assumptions, principles, and causes and without immediate need for statistical techniques, whereas the approach of Brownlee, Greenwood, and co-workers was *a posteriori*, starting from data of past outbreaks (epidemic curves) and working backwards to underlying principles (Kingsland 1985), where statistics is the natural and immediately neces-

sary tool. In short, the difference might be explained by drawing on the difference between Snow's and Farr's hypothesis of epidemic decline. It is a pity that Soper had no opportunity to expand his work, because his ideas clearly aimed at a mix of both approaches.

5.6 | A SCIENCE TAKING FLIGHT

One aspect of old paradigms is that the assumptions underlying the principle tend to be obscure or missing in most work that adopts the paradigm in the early years. These assumptions are not fully realized by the first advocates of a paradigm. The first authors do not know that they are introducing a paradigm or even a metaphor, because for this realization they need a context in which the assumption must be replaced by something more complex. This context needs to grow with applications and with developing theoretical insight. With the context comes exceptions to the rule, and with the exceptions comes a clearer understanding of the underlying principles. An example is the homogeneous assumption as a global descriptor of interaction: any two randomly picked individuals have the same probability per unit of time to come into contact with each other. Systems adhering to mass-action are, in a sense, *well stirred*. Another principle is that the contact rate is proportional to a product of *densities*, that is, numbers of individuals of the various types per unit area. Both assumptions are clear if you bring to mind the origin of mass-action in chemical reactions, where molecules in a well-stirred reaction vessel collide with a rate proportional to the product of the respective concentrations of the various types of molecule. A likely positive exception to the ignorance of assumptions in the case of mass-action is the work of Ross because he did not introduce mass action out of context; Ross introduced mass-action as a natural step in the theory of happenings that he created.

There are also a few *consequences* of mass-action that were realized only slowly. One is that it is linear in the density of susceptibles. This implies that doubling the susceptible density would also double the density of contacts made per unit of time. For many types of contact this is clearly not justified, certainly not over a large range of densities. Another aspect is that in the discrete-time formulation of Hamer and Soper, researchers can overestimate the number of susceptibles that become infected, in the sense that in some time steps more susceptibles can be expected to succumb than are available, thus leading to negative population sizes. Researchers should therefore be careful in formulating discrete-time models with mass-action (see Diekmann & Heesterbeek 2000).

Of these aspects, the first to give rise to new descriptions of interaction is the theory of Wade Hampton Frost (now known unfairly as the Reed-Frost model) from 1928. This model does not suffer from overestimation of new victims, at least not in comparison with the discrete-time mass-action model. Frost never published his model, but the text of a lecture from 1928, in which he describes the idea, was eventually published (as Frost 1976). There is substantial literature on the Frost model and its relations to what is often called the *Soper model* (e.g., see Jacquez 1987, Dietz & Schenzle 1985, and the references given there). The first time these models and the Ross-McKendrick approaches appeared together in a detailed analysis of their differences and similarities was in work by Wilson and Burke (1942, 1943); they later appeared in the papers by Wilson and Worcester (1945). Unknown to all these authors was the work by En'ko in Russian in the late nineteenth century who published, also in connection to measles, another alternative mathematical model for discrete time (described earlier). In contrast to the similar Reed-Frost model, En'ko did not provide a mechanistic basis for his approach. Reed and Frost, however, went so far as constructing a mechanical analogue for their model, where black and white balls were mixed according to certain well-defined rules in a one-dimensional trough (see Fine 1977 for a detailed account).

Mass-action is still in high demand (e.g., it is recognized as making an important contribution to a recent theory for modelling the dynamics of structured populations; Diekmann, Gyllenberg, & Metz 2003). That it is still a relevant paradigm partly stems from the fact that the particular question a researcher wants to study with a model dictates the assumptions and ingredients that will underlie it. When results turn out not to be very sensitive to the precise metaphor used to describe interaction, then it makes sense to stick with the simplest descriptor. This does not imply that it is easy to come to that conclusion or that anybody goes to the effort of testing these assumptions within the model, but certainly not all criticism of mass-action descriptions in present-day models is justified: The nature of the problem studied influences this heavily. Surely Occam's razor principle has a large hand in the popularity of mass action in ecology and epidemiology. Researchers can interpret it as a valid *null model*, for which they realize in many situations that it is not a good mechanistic description of interactions but against which they can test the influence of more detailed mechanisms.

In recent decades, increasingly more involved mechanisms of interaction are required, because various types of heterogeneity—for example, social structure, spatial structure, and age, sex, and behaviour differences—have been shown to influence contact pattern, susceptibil-

ity, and infectivity. Also, more detailed information about these differences is now available and can be analyzed. The differences influence the mechanisms that operate on the level of the individuals and can be important in shaping phenomena at the population level where many such individuals interact. This has given rise to the modelling of a large variety of more heterogeneous types of interaction and more local mixing structures. Some of these models can be put on a strong mechanistic base, but many are heuristic and others lack even that. Especially in the last decade, however, more sophisticated methods have been developed. These methods will be treated in the next chapter by Matt Keeling.

Mass-action stood at the basis of the emerging study of epidemic phenomena as a science. One can wonder whether the many descriptions of the contact process that came later in the development of epidemiology and ecology can ever again have such a far-reaching influence and whether any of these paradigms will remain to influence the way contacts are modelled in these sciences as much as mass-action still does.

ACKNOWLEDGEMENTS

I would like to thank Klaus Dietz for detailed comments on an earlier version of this chapter.

REFERENCES

Aitchison, J., & Watson, G.S. (1988). A not-so-plain tale from the Raj. *In* "The Influence of Scottish Medicine: A Historical Assessment of Its International Impact" (D.A. Dow, Ed.). Parthenon Publishing Group, Carnforth, UK, pp. 113–128.

Bastiansen, O., Ed. (1964). "The Law of Mass Action: A Centenary Volume, 1864–1964." Det Norske Videnskaps-Akademi i Oslo.

Diekmann, O., Gyllenberg, M., & Metz, J.A.J. (2003). Steady-state analysis of structured population models. *Theor. Popul. Biol.* **63**, 309–338.

Diekmann, O., & Heesterbeek, J.A.P. (2000). "Mathematical Epidemiology of Infectious Diseases: model building, analysis and interpretation." John Wiley & Sons, Chichester, UK.

Dietz, K. (1988). The first epidemic model: A historical note on P.D. En'ko. *Austral. J. Stat.* **30A**, 56–65.

Dietz, K., & Heesterbeek, J.A.P. (2005). "Epidemics: The Discovery of Their Dynamics." In preparation.

Dietz, K., & Schenzle, D. (1985). Mathematical models for infectious disease statistics. *In* "A Celebration of Statistics: The ISI Centenary Volume" (A.C. Atkinson & S.E. Fienberg, Eds.). Springer-Verlag, New York, pp. 167–204.

Lund, E.W., & Hassel, O. (1964). Guldberg and Waage and the law of mass action. *In* "The Law of Mass Action: A Centenary Volume, 1864–1964" (O. Bastiansen, Ed.). Det Norske Videnskaps-Akademi i Oslo, pp. 37–46.

Martini, E. (1921). "Berechnungen und Beobachtungen zur Epidemiologie und Bekämpfung der Malaria." Gente, Hamburg.

McKendrick, A.G. (1911). The chemical dynamics of serum reactions. *Proc. Roy. Soc. Lond. B* **83**, 493–497.

McKendrick, A.G. (1912). The rise and fall of epidemics. *Paludism* **1**, 54–66 (*Transactions of the Committee for the Study of Malaria in India*).

McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.* **44**, 98–130.

McKendrick, A.G. (1940). The dynamics of crowd infection. *Edinburgh Med. J.* **47**, 117–136.

Mellor, J.W. (1905). "Higher Mathematics for Students of Chemistry and Physics: With Special Reference to Practical Work," 2nd edition. Longmans, Green & Co., London.

Nye, E.R., & Gibson, M.E. (1997). "Ronald Ross, Malariologist and Polymath: A Biography." Macmillan Press, Houndmills, UK.

Ransome, A. (1880). On epidemic cycles. *Proc. Manchester Lit. Phil. Soc.* **19**, 75–96.

Ross, R. (1911). "The Prevention of Malaria," 2nd edition. John Murray, London.

Ross, R. (1916). An application of the theory of probabilities to the study of *a priori* pathometry: Part I. *Proc. Roy. Soc. Lond. A* **92**, 204–230.

Ross, R., & Hudson, H.P. (1917). An application of the theory of probabilities to the study of *a priori* pathometry. *Proc. Roy. Soc. Lond. A*, **93**, 212–225 (Part II), 225–240 (Part III).

Snow, J. (1853). "On Continuous Molecular Changes, More Particularly in Their Relation to Epidemic Diseases." J. Churchill, London.

Soper, H.E. (1929). The interpretation of periodicity in disease prevalence. *J. Roy. Stat. Soc.* **92**, 34–61 (followed by discussion: 62–73).

Wilson, E.B., & Burke. M.H. (1942). The epidemic curve. *Proc. Natl. Acad. Sci. USA* **28**, 361–367.

Wilson, E.B., & Burke, M.H. (1943). The epidemic curve, Part II. *Proc. Natl. Acad. Sci. USA* **29**, 43–48.

Wilson, E.B., & Worcester, J. (1945). The law of mass action in epidemiology. *Proc. Natl. Acad. Sci. USA* **31**, 24–34 (part I), 109–116 (part II).