REVIEW ARTICLE

A BRIEF HISTORY OF R_0 AND A RECIPE FOR ITS CALCULATION

J.A.P. Heesterbeek

Faculty of Veterinary Medicine, University of Utrecht, Quantitative Veterinary Epidemiology Group, Yalelaan 7, 3584 CL Utrecht, The Netherlands.

Received 27-V-2002

ABSTRACT

In this paper I present the genesis of R_0 in demography, ecology and epidemiology, from embryo to its current adult form. I argue on why it has taken so long for the concept to mature in epidemiology when there were ample opportunities for cross-fertilisation from demography and ecology from where it reached adulthood fifty years earlier. Today, R_0 is a more fully developed adult in epidemiology than in demography. In the final section I give an algorithm for its calculation in heterogeneous populations.

1. INTRODUCTION

The basic reproduction ratio (or number) R_0 is arguably the most important quantity in the study of epidemics and notably in comparing population dynamical effects of control strategies. The quantity is defined as the expected number of new cases of an infection caused by a typical infected individual in a population consisting of susceptibles only. In the last 10-15 years R_0 is an ingredient in almost all papers that use some mathematical modelling in studying the spread of infectious agents.

In demography and ecology, R_0 has the analogous interpretation of the expected number of female offspring born to one female during her entire life. Even though R_0 has eventually pervaded epidemiology much deeper than demography and ecology, the use of R_0 in its present form is of relatively recent origin in epidemiology. In demography the concept necessary for the problems studied there, was basically fully formed in Dublin and Lotka (1925). In contrast, the concept as we know and use it today in epidemiology took until 1980 to become developed and its potency fully realised. This is surprising if we realise that historically there were several opportunities for interaction and cross-over between demography, ecology and epidemiology with Alfred J. Lotka, notably, as a researcher working in all fields.

In this paper I trace the historical development of R_0 in epidemiology and demography/ecology. I hypothesise on how it was possible that the development of what is with hindsight obviously the same concept in two disciplines can have taken so long to converge, even though there was ample opportunity to speed up convergence. I

Acta Biotheoretica **50**: 189–204, 2002.

will present the story more or less chronologically in time, going back and forth between demography and epidemiology whenever necessary.

Thanks to the effort and work of Klaus Dietz, and notably Roy Anderson and Bob May from the late nineteen-seventies to the present day, R_0 has certainly come to stay in epidemiology. Over the last ten years we have seen a fast development with R_0 being applied in increasingly realistic and therefore more complicated situations. Two starting points for that development are the influential book of Anderson and May (1991), on the applied side, and the paper by Diekmann *et al.* (1990), on the mathematical, methodological side. Since then it has become increasingly more clear how to actually compute R_0 in complex situations. I end this paper with an algorithm for its calculation that can be applied to many heterogeneous deterministic situations involving "micro-parasites". A more elaborate version of this "recipe" can be found in Diekmann and Heesterbeek (2000).

The essence of the slow convergence between the demographic and epidemiological concept lies, I postulate, in the different starting points. I can make this clear in the most elementary situation of a population of (spatial) density N with homogeneous mixing where each individual produces on average βN offspring (either in the demographic/ecological or epidemiological sense) per unit of time for an average of $1/\gamma$ time-units. We then have

$$R_0 = \frac{\beta N}{\gamma}.$$

The value of the concept lies in its threshold property: if $R_0 < 1$ the population of infecteds/females cannot grow. In demography this is usually an unwanted situation, whereas in epidemiology this is the situation to strive for using control methods. Another way of expressing the same idea is to rewrite the threshold condition as a relation for the population density N. In the case of epidemiology we then see that in populations with density

$$N < N_c = \frac{\gamma}{\beta}$$

the infection cannot grow. Therefore, instead of thinking in terms of R_0 one can think in terms of a critical population density. (I have chosen not to go into details about how to more realistically formulate contact structure and transmission as a function of population size since current knowledge should not interfere with historical development.)

Historically speaking, what is basically the same threshold concept originated in demography/ecology with the interpretation in terms of reproduction potential (expected numbers of offspring), and in epidemiology in terms of a critical population density. It took a long time for modellers in epidemiology to realise that the formulation in terms of reproduction potential is a much clearer and more powerful concept for infectious diseases as well, which is moreover much more amenable to generalisation to heterogeneous populations, and can be tied much more easily to data and hence applications. An important reason for this long delay in epidemiology (Bob May and Klaus Dietz, personal communication) can indeed be this link to data. The early development of R_0 in ecology/demography had a much closer link to empiricism than the early development in epidemiology in the hands of Kermack and McKendrick

and others, who were much more interested in presenting a mathematically coherent theory. After the realisation, around 1975, that the reproduction potential was to be preferred over critical size, the major hurdle that had to be taken was to tie the formal concept to empiricism. The use of R_0 finally took off when it was found that the quantity could be estimated from readily available data.

2. EARLY IDEAS IN DEMOGRAPHY

In 1886 Richard Böckh of the Registrar's Office in Berlin wrote an appendix to the demographic data for the year 1879 of the city of Berlin (Böckh, 1886). Included in the life-tables in that publication is for the first time a fertility table in which he recorded the age-dependent fertility and survival for females giving birth in 1879. From this he calculated what he called "die totale Fortpflanzung der Bevölkerung" ("the total reproduction of the population") being the sum over all age classes of the product of survival and fertility. This resulted in 2172 births, both males and females, for 1000 females (one can criticise this calculation since all data were derived from a single year). He corrected using the sex ratio and arrived at a first ever estimate of 1.06 for what could be called R_0.

As is the case in many inventions of great concepts, the initial idea of Böckh was not the basis of the further development of R_0. His ideas were followed-up by his student Kuczynski who carried out Böckh's calculations for Berlin for the years 1891-1895 and developed a theory behind the concept in 1932 (Kuczynski, 1932). He did this however, knowing the work of Lotka who had, with Dublin and not in response to Böckh's initial work, fully developed the theory of R_0 by 1925. In the meantime there were several other independent inventions of the same concept in rudimentary form, for example by Knibbs (1917), who called it "crude fertility" or "corrected fertility", when referring only to female offspring.

The development by Alfred Lotka (1880-1949) is interesting as one can get a hint of the thought process he went through over the years, culminating in his "net fertility" in 1925 (1925a), by looking at his successive papers on demography. Lotka worked for the Metropolitan Life Insurance Company in New York and started the chain of reasoning with a short note in *Science* in 1907 on the "rate of natural increase per head", which he called r, of a population with constant birth and death rate. He showed that a fixed age distribution is reached. In 1911 he continues this work in a paper with F.R. Sharpe. They allow the birth rate to be age-dependent, and show that the fixed age distribution is in fact "stable" in the sense that the population composition returns to this distribution after it has been perturbed from it. It is interesting that all quantities are interpreted only for males rather than females. In this paper Lotka for the first time writes the relation

$$r > 0 \Leftrightarrow \int_0^\infty b(a)p(a)da > 1 \qquad (1)$$

but merely as a single step in a line of mathematical reasoning. No special attention is given to the relation or to a possible biological interpretation of the integral. Here $b(a)$ is the rate of giving birth for an individual of age a and p is the probability of survival until age a.

As is often the case in trying to re-create someone's thought process on the basis of published documents alone, one can only guess how Lotka's thoughts on the matter developed. In 1913 he publishes a paper called "a natural population norm", restating the above relation, but this time (apparently having given the integral some thought) adds that "the integral represents the ratio of the total male births in two successive generations". He did not pursue the matter further at that time.

3. MEANWHILE IN EPIDEMIOLOGY

As in demography, epidemiology also had its multiple independent starting points. The most notable of these is in the work of the En'ko who appeared to have a good understanding of threshold phenomena in relation to childhood infections (see En'ko, 1889; Dietz, 1988). I have chosen only to follow the more or less continuous path to bring us to the point in 1919 when Lotka reacted to a series of papers by Ronald Ross and Hilda Hudson and uses relation (1) to settle one of Ross' open problems. Ronald Ross (1857-1932) was a medical doctor, a colonel in the British army, a minor poet and a self-taught mathematician (see Nye and Gibson, 1997 for a recent biography). He led several anti-malaria campaigns, e.g. in Sierra Leone (see below), and discovered in 1898 that (bird) malaria was transmitted by mosquitoes and that malaria was not caused by "bad air" from marshes as was previously believed. He received the Nobel prize for this discovery in 1902.

His work in epidemic modelling started with showing that trying to control malaria by fighting mosquitoes was a real possibility (e.g. Ross, 1911). This in contrast to general opinion at the time that fighting mosquitoes was a difficult route because it would be practically impossible to kill all mosquitoes locally and therefore impossible to stop transmission of malaria. Ross identified the main factors in malaria transmission and calculated the number of new infections arising per month as the product of these factors. Ross referred to his discovery as his "Mosquito Theorem" (for a clear review see Fine, 1975). Essentially, his result is that instead of having to eradicate all mosquitoes in a given area, it suffices to depress the ratio of mosquitoes to man below the threshold, i.e. there exists in Ross' words a "critical density of mosquitoes" below which the malaria parasite cannot be sustained.

Ross clearly thought in terms of a critical threshold and never indicated in his publications that he realised that there would be an alternative way of stating the same condition, i.e. as a reproduction threshold. With hindsight this is not at all surprising. First of all there was at that time no demographic theory that he could borrow the idea from. The second reason may be equally important: it was unfortunate for the development of R_0 that Ross' main interest at the time was malaria, which is a vector-transmitted infection (as Ross himself discovered). While in principle the reproduction concept R_0 can be clearly defined for host/vector systems, they are less straightforward as a starting place than directly horizontally transmitted agents, where (perhaps ironically) the critical threshold description (critical community size) is much more problematic.

Ross published several papers on modelling malaria transmission before he embarked upon constructing a general theory of epidemic phenomena: his "theory of happenings". In a series of three papers (two co-authored with Hilda Hudson, Ross, 1916; Ross and Hudson, 1917a,b), he developed an approach he referred to as "*a*

priori pathometry". Ross was the first to try and develop a general theory of epidemic phenomena using prior assumptions about mechanisms that could be acting in the spread of infections (rather than trying to obtain insight *a posteriori* by studying real epidemics). In addition he and Hilda Hudson did this without mentioning specific infectious agents. In this sense the three combined papers can be called the first developments in abstract (or: modern) epidemic theory.

In the third paper there appears for the first time a very general equation to describe the population dynamics of an infectious agent. For the equation Ross and Hudson introduce the time s elapsed since an individual became infected. The equation then gives the incidence at time t

$$F_{t,0} = \frac{A}{P}\int_0^\infty cF_{t,s}ds \tag{2}$$

where c is the infectivity and where s is infection-age (i.e. time elapsed since infection); $F_{t,s}$ is the number of cases that at time t have infection-age s, A is the unaffected population and P is the total population size. Although the equation is formulated in this way, Ross and Hudson only regard the special case where the infectious period is constant (i.e. $F_{t,s}$ is positive constant for $q_1 < s < q_2$ and 0 for other values of s).

Ross and Hudson concentrated on directly transmitted infections however, and this might (in light of what was discussed above) be the reason that they did not try and generalise the critical density concept to general infectious agents. The whole idea does not come up at all. Ross does not make the connection to his groundbreaking malaria insight. We will see later that it is the above equation though that would trigger Anderson McKendrick and W.O. Kermack to make up for Ross' omission.

This ends Ross' involvement in the genesis of R_0. In summary, we can say that there are two main reasons for the slow development of the concept in epidemic theory. Firstly, the starting point as a critical density rather than a reproduction threshold idea, and secondly the fact that Ross developed the idea specifically for the vector-transmitted malaria, (where looking for a critical density of mosquitoes seems a natural approach).

4. LOTKA IN EPIDEMICS AND MATURATION IN DEMOGRAPHY

The failure of Ross to translate his malaria critical density to his theory of happenings is all the more a pity because Lotka studied the 1917 papers of Ross and Hudson in detail and was fascinated by it. He published a reaction in 1919 called "A contribution to quantitative epidemiology". He assumes a constant population of susceptibles (and so is concentrating on the early phase of an epidemic), but allows the infectivity to be a function of infection-age $c = c(s)$, and does not impose Ross' conditions on F. He shows that the integral equation (2) then has solutions for the total number of cases at time t of the form $z(t) = z(0)\exp(ut)$, where u is the only real root of the associated characteristic equation and is positive when the ingredients satisfy

$$u > 0 \Leftrightarrow K\int_0^\infty p(s)c(s)ds > 1. \tag{3}$$

Here $K = A/P$ and p is the survival probability once an individual is infected. We immediately note the similarity with relation (1) of Sharpe and Lotka in their 1911 paper on the rate of natural increase r. Lotka noted the similarity in broad terms as well as remarking that "growth of population and spread of a disease" are very similar from a mathematical point of view. The "mathematical" similarity of method is all he notes because he does not transpose the demographic interpretation he gave to his integral inequality in 1913 into epidemiological terms. For epidemiology, R_0 could have started here, but it did not.

Lotka returned to infectious diseases in 1923 in a series of long papers devoted to malaria, but there he almost drowns in system theory, which was a topic very much on his mind at the time. He does mention thresholds for persistence of malaria, but phrases this in terms of stability of the malaria-free steady state with a quantity being less than or equal to 0. This is natural since he is analysing systems describing spread in real time. He does not make the connection to his demographic analysis on a generation basis.

In the demographic development of his ideas Lotka had identified the integral first as merely a step in a mathematical proof and then in 1913 as an integral having an interpretation. Finally in the early nineteen-twenties the realisation came that the integral was an interesting quantity in its own right. Lotka is working on his 1925 book *Elements of Physical Biology* (1925b). He treats his work on Ross' malaria equations as an example of his more general theory of systems analysis in biology. In a subsequent section of the book he treats the work of William Robin Thompson. Thompson, like Lotka, plays an important role in the birth of theoretical ecology (see Kingsland, 1985, who also lists several of Thompson's publications; see also Smith and Keyfitz, 1977, who reprint Thompson's paper on reproduction with overlapping generations, with contributions by H.E. Soper). He was an entomologist interested in host-parasitoid relations. In his work he defines "the reproductive power" of the parasite (see e.g. Thompson, 1923), which is the number of eggs deposited per female. Since he assumes that one egg only is deposited per host this is essentially R_0. Now Lotka recognises that the "rate of multiplication per generation" is closely related to the ratio of the total births in successive generations, which he denotes by a letter for the first time

$$R = \int_0^\infty b(a)p(a)da. \tag{4}$$

Lotka does not give it much more attention than this in his book. It must therefore have been Dublin's influence that in their seminal paper also from 1925 Lotka finally crosses the threshold. The paper is titled "On the true rate of natural increase" and on page 310 they write, following a calculation with a fertility table:

> "The net result is that if we follow the history of 100,000 females at the current rate of fecundity we find that throughout their life they give birth to 116,760 daughters; or, on an average, one female gives birth to 1.168 daughters in the course of her life. This, then, is the ratio of the total births of

daughters in two successive generations. It will be convenient for future reference to denote this ratio by the symbol R_0."

The authors do not attach a name to R_0 in this paper. Lotka will later refer to it as "net fertility" and still later as "net reproductive rate", probably as a result of reading Robert Kuczynski's 1932 and 1935 books or one of his earlier works. (Incidentally, Lotka and Kuczynski knew each other's work well and had frequent arguments in publications, notably also about who had scientific priority in various matters, among them the reproduction threshold; see Samuelson, 1977, for a fascinating account.) Since initially Lotka does not use the term involving "reproduction", but does from the very first idea think in terms of "ratio in successive generations", it seems most likely that he chose the symbol R to signify "ratio" rather than "reproduction". The subscript '0' (absent from Lotka's book of the same period) arises in the Dublin-Lotka paper from a Taylor expansion of the mean age at childbirth, given in their appendix, where R_n is defined as $\int a^n p(a)m(a)da$ (where p is again age-dependent survival, and where m is the age-dependent fertility). The analysis of Dublin and Lotka and parts of Lotka's earlier work were later given mathematical rigour by Feller (1941).

The "true rate of natural increase" from the title of the Dublin-Lotka paper is not R_0, but r. In demography the authors feel this quantity, rather than R_0, is the more natural measure of population growth. Although several authors, notably Feller, would take demographic population dynamics further and provide a more rigorous mathematical basis for it than Lotka did, still the development of R_0 in demography basically ends here in 1925. It will take a further 50 years for the development in epidemiology to reach the same level. In contrast to demography though, the concept of R_0 is very powerful as it is directly related to the amount of control effort needed to eliminate an infection from a population. This difference in use makes that the concept in epidemiology ultimately called for a substantially more detailed development beyond the final halting point in demography. For further details in demography and ecology (e.g. the approximation $R_0 \approx \ln(r)/T$, where T is the generation time; see May, 1976 for a clear exposition, and also Dublin and Lotka, 1925) see the original papers and treatment in Smith and Keyfitz (1977), Kingsland (1985) and the more modern treatment in e.g. Begon *et al.* (1998).

5. EPIDEMIOLOGY ENTERS THE MODERN AGE

Anderson Gray McKendrick was a medical doctor who served in the British army and in that capacity served under the command of Ross in Sierra Leone in 1901 during one of the anti-malaria campaigns (Aitchison and Watson, 1988). Ross stimulated the young McKendrick to learn to apply mathematical reasoning to medical problems. This was most prominent during their journey by boat back to England in 1901. In later correspondence between the two men, McKendrick reports on his progress in the self-study of mathematics. Ross in his correspondence with McKendrick (Ross Archives, LSHTM) makes clear that he is interested in establishing a general theory of epidemic phenomena ("happenings"), he wanted "to establish the general law of epidemics". He recognised McKendrick as someone who would be able to carry his work further. 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."

McKendrick published several papers on mathematical topics in medicine (among them the first paper containing an age-structured PDE-model in 1914), before he met W.O. Kermack. At that time McKendrick was Superintendent at the Royal College of Physicians Laboratory in Edinburgh. Kermack was one of his employees; he became blind when a chemical experiment went badly wrong. He and McKendrick published their first, and important, joint paper in 1927 (Kermack and McKendrick, 1927), and many were to follow (not only on epidemics, but also on purely mathematical topics such as properties of points arranged on a Mobius surface; see Harvey, 1943 for an obituary of McKendrick and an almost complete list of publications).

In their 1927 classic paper, which is arguably the most misquoted paper in epidemic theory, they systematically follow-up earlier work by McKendrick and notably the papers by Ross and Hudson. Where Ross failed to translate his critical density idea to other infections than malaria, Kermack and McKendrick more than make up for this omission. They prove their celebrated threshold theorem which states that, in order for an infectious agent to be sustained in a population, the population density N has to exceed a certain critical density $N_c = 1/A$, where

$$A = \int_0^\infty \phi_t B_t dt.$$

In this formula, B_t is the probability that a newly infected individual is still infected at infection-age t, and ϕ_t is the infectivity at infection-age t. Note that Kermack and McKendrick generalise Ross' critical density concept and do not think in terms of a reproduction quantity with threshold 1; the reproduction ratio would be NA in their notation.

The reproduction formulation finally came 25 years later when, in 1952, George Macdonald published a paper "The analysis of equilibrium in malaria" in the *Tropical Diseases Bulletin*. Macdonald was the director of the Ross Institute at the London School of Hygiene and Tropical Medicine. He devotes his crucial paper entirely to malaria, but takes a more general view of epidemic phenomena in one paragraph of his appendix:

> "Basic Reproduction Rate of malaria, Definition. The number of infections distributed in a community as the direct result of the presence in it of a single primary non-immune case."

Macdonald refers to the work by Ross and Kermack and McKendrick as the "theory of critical level". He gives an expression for his basic reproduction rate for malaria in terms of the ingredients identified by Ross and then explains that by rewriting the resulting expression he can obtain the critical level derived by Ross. In all probability, Macdonald was the first to do this in epidemiology and he should also be credited with introducing the name "basic reproduction rate". Macdonald applied his ideas as an advisor to the World Health Organisation (WHO) as a basis of anti-malaria campaigns. In 1955 Macdonald starts using the symbol z_0 for his quantity. He does not use a symbol involving R and also does not mention any of the work done in demography. We are therefore led to conclude that Macdonald did not see the connection to demography, just as Lotka had not realised the connection to epidemiology (although to be fair to Macdonald, this is more unfortunate for Lotka since he worked in both areas). Reading demographic literature could certainly have

prevented Macdonald from using "rate" instead of "ratio". Since R_0 is a dimensionless quantity, "ratio" is the more accurate suffix. Unfortunately, the word "rate" has persevered in epidemiology for a long time.

6. THE MATHEMATICIANS TAKE OVER

One would expect that with the attention Macdonald gave to his quantity the field would surely be on par with demography soon since in the early nineteen-fifties a large number of mathematicians entered the field of epidemic spread. There was a setback however. The setback was a very important event in the development of modern epidemic theory, but alas at the same time a development that slowed down the genesis of R_0. Norman Bailey published in 1957 the first book entirely devoted to the mathematical study of epidemic phenomena. Notwithstanding the many qualities of this work, it shows the dangers of reviewing the state of the art: when the reviewer misses a potentially important development it may get lost for years. This was the case with Bailey's book. He recognised many developments and opened up the subject for mathematicians who not only found a good introduction to epidemics, but also a wealth of mathematical problems. Maurice Bartlett, Peter Whittle, David Kendall and many others read Bailey's book, and took it as their main source of information; certainly Bartlett also read several of Bailey's earlier papers. They all knew the work by Kermack and McKendrick, but as mathematicians would not easily be led to read a paper in the *Tropical Diseases Bulletin* unless they would be told that it contained a mathematically interesting idea.

Bailey did read the paper by Macdonald, and quotes it in his book. He apparently did not recognise the potential of the definition in the appendix of the "basic reproduction rate for malaria" for a much more general class of infections; he does not mention the idea at all in his book. It is no wonder therefore that none of the mathematicians was enticed to read the original Macdonald papers for a number of years to come.

Bailey did elaborate on a threshold concept in his book (Bailey, 1957) that he had introduced in a paper in 1953. There he defined a quantity ρ, which he called "the relative removal rate". In terms of the ingredients given in the Introduction above, $\rho = \gamma/\beta$, in other words, Bailey's relative removal rate is equal to the critical density. He recognises the threshold connection and remarks that, for the introduction of the infectious agent into a susceptible population of density x_0, we expect an epidemic if $\rho < x_0$, and no epidemic if $\rho > x_0$.

To show how rapidly things could have developed had Bailey picked up on Macdonald's idea, I quote from Bartlett's 1955 book on stochastic processes (which quotes Bailey's papers). Bartlett treats several examples and two of them are devoted to population growth and epidemic models, Sections 4.3 and 4.4, respectively. In the section on demographic theory he explains the theory "R, the net reproduction rate defined as the mean of the female replacement distribution per female". In the next section however, he explains the basic epidemiological counterpart in terms of the critical threshold of susceptibles and not in terms of reproduction potential.

While Bartlett's book dealt with stochastic processes he was not the one to extend the critical density concept to the stochastic situations with a finite number n of susceptible individuals at the moment of introduction of an infectious agent. This was

done by Whittle (1955). For the simple case, the Kermack and McKendrick ordinary differential equation model, he showed that for $\rho_n < n$ (note the similarity with Bailey's threshold theory) we will get an epidemic with probability $1 - [\rho_n/n]^a$, where a is the initial number of infectives introduced into the population. Even though this is the form in which this result is always quoted, Whittle actually proved a more subtle result. He first specified, by means of an additional parameter, the pre-determined fraction of the susceptibles that has to become infected before we actually speak of an epidemic (which might of course depend on population and parasite characteristics). Bailey in his 1957 book calls this fraction the "intensity". Whittle's original theorem includes this parameter and the result quoted above is obtained when this parameter approaches zero (i.e. when any fraction of the susceptibles that becomes infected is interpreted as constituting an epidemic). The stochastic threshold theory was extended by Barucha-Reid (1956) among others (see Dietz (1995) for a short list).

For a number of years the theory of epidemics blossomed, but by the end of the nineteen-sixties much of the original momentum had waned. The field came certainly no closer to defining R_0. In 1974 a conference was held on epidemiology in Utah. Many young researchers, relatively new to the field, were present. Two people re-introduced/reinvented the concept (Hethcote, 1975, Dietz, 1975). The stochastic development also came closer to applications in 1975 in a paper of Niels Becker. Herb Hethcote introduced the "infectious contact number", but it was Klaus Dietz who for the first time clearly defines:

> "The quantity R is called the reproduction rate, since it represents the number of secondary cases that one case can produce if introduced to a susceptible population."

Dietz moreover shows that in the case of a mosquito-born agent the critical threshold condition translates into $R > 1$ or $R < 1$. He also indicates the relation to Bailey's relative removal rate. It took almost 20 years after the book by Bailey for a mathematician to go back to the original papers and recognise the value of the concept. Bailey worked at WHO from 1966 and there he hired Dietz to work on malaria. Dietz therefore was very familiar with Macdonald's contributions. In fact, just one year earlier in Dietz *et al.* (1974) concerning malaria, he uses Macdonald's name *basic reproduction rate* and also his symbol z_0. In between these papers, Dietz must have discovered the publications of C.E.G. Smith on arboviruses. Smith (1964) uses the symbol *RR* (and quotes a PhD-thesis of Macnamara, 1955, from Cambridge as the originator). Dietz quotes and credits both, decides that a symbol involving R is more sensible (as Lotka had discovered long ago) than Macdonald's z_0, and removes the superfluous and inconvenient second R. Smith was the first to derive the well-known relation between R_0 and the fraction of a population to vaccinate.

Apart from clearly defining R_0 as a mathematical concept, it is a major step forward that Dietz also showed in his 1975 paper that R_0 could be related to data quite easily as the inverse of the proportion of susceptibles in an endemic steady state (and as the ratio of life expectancy and average age at infection).

7. AT LAST: THE ECOLOGISTS

Finally, one could say, a clear definition of R_0 as a reproduction concept was given in epidemiology and surely now the use of the concept in looking at control of

infectious disease agents with epidemiological models would start to grow. Notably Bob May and Roy Anderson were the ones to successfully advocate the use and value of R_0. Even though May and Anderson would turn out to be the most vigorous advocates in the early nineteen-eighties, it took a number of years for them to realise the potential. Both came to epidemiology from a more ecological, data-driven background rather than a mathematical background, and this certainly had an impact on their approach. In 1979 (Anderson and May, 1979a,b), they publish their very influential two-part *Nature*-paper on the population biology of infectious diseases. These papers have played a dominant role in revitalising the subject of infectious disease modelling, after attention for it had waned from the late nineteen-sixties. After the SIMS proceedings were published four years earlier, this would have been the perfect opportunity to popularise R_0. The papers are sometimes credited with indeed introducing R_0 into modern epidemic theory. However, the symbol does not appear in the two papers, in fact the reproduction idea is not even mentioned: the whole analysis is done in terms of critical sizes of host populations. Ironically, the paper by Dietz from 1975 and the earlier work by Macdonald (both containing the reproduction interpretation, and Dietz containing the necessary link to data), as well as the paper by Yorke *et al.* (1979) (who write R and call this quantity "transmissibility"), are quoted by Anderson and May, albeit in different contexts.

It is in the short time frame from 1979-1980, that Anderson and May came to realise the full potential of R_0. For example, Anderson and May (1982a) publish a paper in *Science* in February of 1982 in which they more than make up for their omission in the *Nature* papers and extensively use R_0. They call it "the intrinsic reproduction rate". In a chapter called "Population ecology of infectious diseases" (contributed to a book on theoretical ecology edited by May in 1981) Anderson uses R instead of R_0. He gives credit to Dietz (and uses his notation) and Macdonald and introduces the name *basic reproductive rate*. But in the first chapter of the same book, May discusses models for population growth and there he explains Lotka's original demographic R_0 as: "the expected number of offspring (or females) produced by the average individual (or female) over its lifetime". The chapters do not refer to each other where this concept is concerned. This shows that, even though the concept has now reached the same stage as in demography 50 years earlier, there still is not the full realisation that the demographic/ecological and the epidemiological concepts are actually identical and there is no consensus, even for the writers themselves, about how to denote and to name the quantity. For their more data-oriented approach it was essential that they saw that R_0 was more than a mere mathematical construct. Dietz had already shown a connection to data in 1975. A paper by Paul Fine and J.A. Clarkson (1982) on measles, which May read in draft in early 1980, also served to bring home the idea that R_0 could play a central role for epidemiology and could moreover be estimated from serological data (May, personal communication).

Once this realisation came, Anderson and May wasted little time. They made promoting R_0 and notably its application in epidemiology one of the major forces behind their famous Dahlem conference in 1982. This conference was to be most influential in reviving scientific interest in applying mathematical modelling as a tool in studying the spread and control of infectious agents. Anderson's name *basic reproductive rate* and Lotka's symbol R_0 are linked and in the proceedings volume of the meeting (Anderson and May, 1982b) the quantity finally gets the attention it

deserves. Strangely enough almost every contribution to the volume uses R_0 heavily and in such a matter-of-fact way that a reader coming to this for the first time will be inclined to conclude that this concept had been well-established, studied and used in epidemiology for decades. We have seen that this is far from the truth.

8. A RECIPE TO CALCULATE R_0

The major developments after 1982 are applications of R_0, links to data and the extension of the original concept to heterogeneous populations. From the early nineteen-eighties heterogeneity in populations played an increasingly dominant role in the problems facing epidemiology. This was stimulated by the onset of the HIV epidemic where clearly analyses that treated all susceptible or infected individuals alike could insufficiently capture the dynamics of spread. Individuals differ in characteristics that are epidemiologically relevant. For example, traits such as age, sex, genetic composition and many others can influence an individual's susceptibility, the patterns of contact to other individuals and the individual's infectivity

The increased awareness that individual differences were very important for understanding the spread and control of infections stimulated epidemic modelling enormously to expand in the direction of higher dimensional systems where these individual characteristics could be taken into account. This also raised the problem of defining and calculating R_0 for heterogeneous populations. Many people contributed to increased understanding of how to calculate R_0 in heterogeneous populations. I do not go into details about the developments in the 20 years since the Dahlem conference proceedings appeared. Notably Anderson and May in their influential book from 1991 devoted a lot of attention to R_0, its use, its estimation from serological and other data and its characterisation in many heterogeneous populations (e.g. discrete structure in many forms, age structure) building on their own work and that of others. Given their main focus on applications and relation to data it is not strange that they do not aim in their book to build a mathematical theory of R_0 in heterogeneous populations. A general methodology was developed in Diekmann *et al*. (1990) and Heesterbeek (1992). This was stimulated by questions and discussions about how to mathematically characterise R_0 in heterogeneous populations in general at a meeting of epidemic modellers at the mathematical institute in Oberwolfach in 1989. Dietz (1993) presents an overview of methods to estimate R_0.

In Diekmann *et al.* (1990) it is shown that to define R_0 we first specify a linear positive operator which we christened "the next-generation operator". This operator maps generations of infected individuals into each other, as distributed over the possible individual characteristics, in the situation where a very small number of infecteds enter a very large population of susceptibles. The susceptibles are assumed to be in a demographic steady state in the absence of the infectious agent. The 'generations' are completely analogous to generations in the demographic sense. The difference is that "being born" is now interpreted as "becoming infected". The fact that the operator is linear stems from the assumption that we regard only the initial stage of an invasion by an infectious agent into a fixed population of susceptibles. We disregard in the initial stage the fact that the susceptible population decreases as a result of the infection. Iteration of this operator with a given initial distribution of infecteds over the characteristics tells us what will happen to this initial generation.

The dominant eigenvalue of the next-generation operator then tells us whether the infected population will grow or decline in size, independent of the initial distribution. This eigenvalue has precisely the desired properties in that it has threshold value 1 and it has the interpretation of expected number of new cases produced per (typical) infected individual (i.e. the average contribution to the next generation). The dominant eigenvalue of the next-generation operator is therefore R_0.

A main point in the definition of R_0 is to think in terms of generations and not in terms of real-time change in a population. Whereas for demography the real-time growth is a more important focus (Lotka's r), in epidemiology the generation growth and R_0 dominate the applications. In part this is because of the interpretation of R_0, which has an intuitively appealing direct link to intervention, eradication and control. Another factor is that calculating r leads to implicit equations, whereas R_0 has an explicit definition. This is best illustrated in the age-structured case, resulting in the famous Euler-Lotka characteristic equation relating r to fecundity b and survival p. Then $R_0 > 1$ if and only if there is a unique $r > 0$ such that

$$1 = \int_0^\infty e^{-ra} b(a) p(a) da$$

with R_0 given by equation (4) (see Feller, 1941 for the first rigorous exposition; see also Heesterbeek and Dietz, 1996). The advantages of this are most prominent in epidemiological situations, where the heterogeneity considered is often more complicated than the mostly age-structured situations in demography. Both aspects mentioned might account for the fact that in demography r became favoured above R_0 whereas in epidemiology it is the other way around. As mentioned earlier, the generation time and r provide an estimate of R_0.

The most frequently encountered problem in the literature when trying to calculate invasion thresholds for a given epidemic model is that one fails to arrive at a quantity with a clear biological interpretation. This arises from taking an unfortunate starting point. One often starts from a real-time description and tries to derive an essentially generation-time quantity from that. One attempts this by manipulating a full system of differential equations and performing stability analysis around the infection-free steady state. While this can certainly lead to threshold quantities, one should keep in mind that these are, as a rule, not equal to R_0. The only thing in common is their predictive behaviour in the neighbourhood of the threshold 1, but generally these quantities have no biological interpretation.

I do not go into more details, but refer to Diekmann and Heesterbeek (2000), where the most elaborate exposition of the mathematical theory can be found. I end this paper with a recipe for the calculation of R_0 that can be applied to many heterogeneous deterministic situations involving "micro-parasites". The original of this recipe, with many examples, can be found in the book quoted above.

Basically there are three steps: identifying the relevant heterogeneous characteristics, constructing the elements of the next-generation operator in terms of basic parameters and ingredients, and computation of the dominant eigenvalue of the operator.

The first step is important. Not all characteristics are relevant; the only relevant characteristics are those that distinguish between individuals at the moment they are

born (in the epidemiological sense). We only regard infected individuals, and the state of the individual at birth (called the "type-at-birth") can depend not only on physiological, genetic or behavioural characteristics, but can also depend on the transmission route via which the individual becomes infected. Potential ingredients of the type at birth are all such characteristics that cause individuals to have different susceptibility, a different contact structure compared to other individuals and different infectivity or reaction to the infection. For example, a vector-transmitted infection with one host species and two vector species, without taking any other form of heterogeneity into account, will have three types-at-birth: individuals can become infected either as a host, as vector 1 or as vector 2. In the case of a discrete number of types-at-birth, the next-generation operator is an n*n-matrix, where n is the number of types-at-birth. To calculate R_0 in the above example we will end up with a 3*3-matrix. When starting from the real-time system, one would start from a nine-dimensional system of differential equations for susceptible, infected and removed individuals of host and vectors and derive a threshold quantity using stability analysis.

Let the types-at-birth be denoted by ξ and η taking values in some space $\Omega \subset \Re^n$.

For the second step, define the element $k(\xi, \eta)$ as the expected number or new cases with type-at-birth ξ caused by a single infected individual with type-at-birth η. The modelling part involved in calculating R_0 consists of deriving expressions for these elements in terms of basic parameters and ingredients. Often one would want to distinguish infecteds also in terms of infection-age τ (which is then integrated out), since infectivity, contacts and disease status are usually dependent on time elapsed since becoming infected. The ingredients then include: infectivity as a function of τ and η, survival as a function of τ and η, and the total contacts towards susceptibles that could be born with type ξ. Now define the next-generation operator K as

$$(K\psi)(\xi) = \int_{\Omega} k(\xi, \eta)\psi(\eta)d\eta$$

where ψ denotes a generation of infecteds as distributed over the types-at-birth.

Finally, R_0 is the dominant eigenvalue of K.

ACKNOWLEDGEMENTS

This paper has benefited greatly from the recollections and insights of Bob May and Klaus Dietz. I thank both of them for taking the time to share their thoughts so generously with me.

REFERENCES

Aitchison, J. and G.S. Watson (1988). A not-so-plain tale from the Raj. In: The Influence of Scottish Medicine. D.A. Dow (ed.). Parthenon, Carnforth, pp 113-128.

Anderson, R.M. (1981). Population ecology of infectious diseases. In: Theoretical Ecology. R.M. May (ed.). Blackwell, Oxford. pp. 318-355.

Anderson, R.M. and R.M. May (1979a). Population biology of infectious diseases. Part I. Nature 280: 361-367.

Anderson, R.M. and R.M. May (1979b). Population biology of infectious diseases. Part II. Nature 280: 455-461.

Anderson, R.M. and R.M. May (1982a). Directly transmitted infectious diseases: control by vaccination. Science 215: 1053-1060.

Anderson, R.M. and R.M. May (eds.) (1982b). Population Biology of Infectious Diseases. Springer-Verlag, Berlin.

Anderson, R.M. and R.M. May (1991). Infectious Diseases of Humans: transmission and control. Oxford University Press, Oxford.

Bailey, N.J.T. (1953). The total size of a general stochastic epidemic. Biometrika 40: 177-185.

Bailey, N.J.T. (1957). Mathematical Theory of Epidemics. Griffin, London.

Bartlett, M.S. (1955). An Introduction to Stochastic Processes. Cambridge University Press, Cambridge.

Barucha-Reid, A.T. (1956). On the stochastic theory of epidemics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Volume IV: Biology and Problems of Health, J. Neyman (ed.). University of California Press, Berkeley. pp. 111-119.

Becker, N. (1975). The use of mathematical models in determining vaccination policies. Bulletin of the International Statistics Institute 46: 478-490.

Begon, M., J. L. Harper and C. R. Townsend (1998). Ecology: Individuals, Populations and Communities. 3rd edition. Blackwell Science, Oxford.

Böckh, R. (1886). Statistsiches Jahrbuch der Stadt Berlin, Volume 12, Statistik des Jahres 1884. P. Stankiewicz, Berlin.

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

Diekmann, O., J.A.P. Heesterbeek and J.A.J. Metz (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology 28: 365-382.

Dietz, K. (1975). Transmission and control of arboviruses. In: Epidemiology, D. Ludwig and K.L. Cooke (eds.). SIAM, Philadelphia. pp. 104-121.

Dietz, K. (1988). The first epidemic model: a historical note on P.D. En'ko. Australian Journal of Statistics 30A: 56-65.

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. Statistical Methods in Medical Research 2: 23-41.

Dietz, K. (1995). Some problems in the theory of infectious disease transmission and control. In: Epidemic Models: their Structure and Relation to Data, D. Mollison (ed.). Cambridge University Press, Cambridge, pp. 3-16.

Dietz, K., L. Molineaux and A. Thomas (1974). A malaria model tested in the African Savannah. Bulletin of the World Health Organization 50: 347-357.

Dublin, L.I. and A.J. Lotka (1925). On the true rate of natural increase. Journal of the American Statistical Association 150: 305-339.

En'ko, P.D. (1889). On the course of epidemics of some infectious diseases. Vrach. St. Petersburg, X, 1008-1010, 1039-1042, 1061-1063 (in Russian). English translation by K. Dietz, 1989. International Journal of Epidemiology 18: 749-755.

Feller, W. (1941). On the integral equation of renewal theory. Annals of Mathematical Statistics 12: 243-267.

Fine, P.E.M. (1975). Ross's a priori pathometry – a perspective. Proceeding of the Royal Society of Medicine 68: 547-551.

Fine, P.E.M. and J.A. Clarkson (1982). Measles in England and Wales. International Journal of Epidemiology 11: 5-14 (part I), 15-25 (part II).

Harvey, W.F. (1943). Anderson Gray McKendrick. Edinburgh Medical Journal 50: 500-506.

Heesterbeek, J.A.P. (1992). R_0. PhD Thesis. University of Leiden.

Heesterbeek, J.A.P. and K. Dietz (1996). The concept of R_0 in epidemic theory. Statistica Neerlandica 50: 89-110.

Hethcote, H.W. (1975). Mathematical models for the spread of infectious diseases. In: Epidemiology, D. Ludwig and K.L. Cooke (eds.). SIAM, Philadelphia. pp. 122-131.

Kermack, W.O. and A.G. McKendrick (1927). Contributions to the mathematical theory of epidemics - I. Proceedings of the Royal Society of Medicine 115A: 700-721 (reprinted in Bulletin of Mathematical Biology 53: 33-55, 1991).

Kingsland, S.E. (1985). Modeling Nature: episodes in the history of population ecology. University of Chicago Press, Chicago.

Knibbs, G.H. (1917). Mathematical theory of population, of its character and fluctuations, and of the factors which influence them. Appendix to Census of the Commonwealth of Australia. McCarron, Bird & Co, Melbourne.

Kuczynski, R.R. (1932). Fertility and Reproduction. Falcon Press, New York.

Kuczynski, R.R. (1935). The Measurement of Population Growth. Sidgwick and Jackson, London.

Lotka, A.J. (1907). Relation between birth rates and death rates. Science 26: 21-22.

Lotka, A.J. (1913). A natural population norm. J. Washington. Acad. Science 3: 241-293.

Lotka, A.J. (1919). A contribution to quantitative epidemiology. Journal of the Washington Academy of Science 9: 73-77.

Lotka, A.J. (1923). Contribution to the analysis of malaria epidemiology. I : General part. Supplement to American Journal of Hygiene 3: 1-37 (parts II to V in same volume).

Lotka, A.J. (1925a). The measure of net fertility. Journal of the Washington Academy of Science 15: 469-472.

Lotka, A.J. (1925b). Elements of Physical Biology. Williams and Wilkins Company, Baltimore.

Macdonald, G. (1952). The analysis of equilibrium in malaria. Tropical Diseases Bulletin 49: 813-829.

Macnamara, F.N. (1955). Man as the host of yellow fever virus. Dissertation for M.D. degree, University of Cambridge.

May, R.M. (1976). Estimating r: a pedagogical note. American Naturalist 110: 469-499.

May, R.M. (1981). Models for single populations. In: Theoretical Ecology. R.M. May (ed.). Blackwell, Oxford. pp. 30-52.

Nye, E.R. and M.E. Gibson (1997). Ronald Ross, Malariologist and Polymath, a Biography. Macmillan Press, London.

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. Proceedings of the Royal Society London A 42: 204-230.

Ross, R. and H.P. Hudson (1917a). An application of the theory of probabilities to the study of a priori pathometry – Part II. Proceedings of the Royal Society London A 43: 212-225.

Ross, R. and H.P. Hudson (1917b). An application of the theory of probabilities to the study of a priori pathometry – Part III. Proceedings of the Royal Society London A 43: 225-240.

Samuelson, P.A. (1977). Resolving a historical confusion in population analysis. In: Mathematical Demography, selected papers, D. Smith and N. Keyfitz (eds.). Springer-Verlag, Berlin. pp. 109-130.

Sharpe, F.R. and A.J. Lotka (1911). A problem in age distribution. Philosophical Magazine 21: 435-438.

Smith, C.E.G. (1964). Factors in the transmission of virus infections from animal to man. Scientific Basis of Medicine Annual Review 1964. pp. 125-150.

Smith, D. and N. Keyfitz (eds.) (1977). Mathematical Demography, selected papers. Springer-Verlag, Berlin.

Thompson, W.R. (1923). La théorie mahématique de l'action des parasites entomophages. Revue Générale des Sciences Pures et Appliqueés 34: 202-210.

Whittle, P. (1955). The outcome of a stochastic epidemic – a note on Bailey's paper. Biometrica 42: 116-122.

Yorke, J.A., N. Nathanson, G. Pianigiani and J. Martin (1979). Seasonality and the requirements for perpetuation and eradication of viruses in populations. American Journal of Epidemiology 109: 103-123.