Epidemiologic Reviews
Vol. 15, No. 2

Printed in U.S.A.

Herd Immunity: History, Theory, Practice

Paul E. M. Fine

INTRODUCTION

Herd immunity has to do with the protection of populations from infection which is brought about by the presence of immune individuals. The concept has a special aura, in its implication of an extension of the protection imparted by an immunization program beyond vaccinated to unvaccinated individuals and in its apparent provision of a means to eliminate totally some infectious diseases. It is a recurrent theme in the medical literature and has been discussed frequently during the past decade. This new popularity comes as a consequence of several recent major achievements of vaccination programs, i.e.: the historic success of the global smallpox eradication program; dramatic increases in vaccination coverage stimulated by national programs and by the Expanded Programme on Immunization; the commitment of several countries to eradicate measles; and international dedication to eliminate neonatal tetanus and to eradicate poliomyelitis from the world by the year 2000.[1]

Received for publication January 27, 1993, and in final form July 29, 1993.

From the Communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1, England. (Reprint requests to Dr. Paul E. M. Fine at this address.)

[1] Though the words "eradicate" and "eliminate" have been used interchangeably by some authors in the past, current usage of *eradication* implies reduction of both infection and disease to zero whereas *elimination* implies either regional eradication, or reduction of disease incidence to some tolerably low level, or else reduction of disease to zero without total removal of the infectious agent (1). Thus the 42nd World Health Assembly recommended "elimination of neonatal tetanus by 1995 and global eradication of poliomyelitis by the year 2000" (2).

Along with the growth of interest in herd immunity, there has been a proliferation of views of what it means or even of whether it exists at all. Several authors have written of data on measles which "challenge" the principle of herd immunity (3–5) and others cite widely divergent estimates (from 70 to 95 percent) of the magnitude of the herd immunity threshold required for measles eradication (6–8). Still other authors have commented on the failure or "absence" of herd immunity against rubella (9, 10) and diphtheria (11). Authorities continue to argue over the extent to which different types of polio vaccine *can*, let alone do, induce herd immunity (12–14). Given such differences of opinion, there is need for clarification.

Many authors have based their discussions of herd immunity on an influential paper published in 1971 by Fox et al. titled "Herd immunity: basic concept and relevance to public health immunization practices" (15). This paper took as its starting point a medical dictionary's definition of herd immunity as "the resistance of a group to attack by a disease to which a large proportion of the members are immune, thus lessening the likelihood of a patient with a disease coming into contact with a susceptible individual" (16). While useful, even this definition lends itself to different interpretations; these may be either quantitative (herd immunity as partial resistance, reflected in reductions in frequency of disease due to reductions in numbers of source cases and of susceptibles) or qualitative (herd immunity as total resistance, implying a threshold number or percentage of immunes above which an infection cannot persist). Each of these interpretations has its place, but they are sometimes confused in debates

on the subject. A given population may exhibit one (partial, quantitative) without the other (total, qualitative) form of herd immunity. It will be found that such definitions do not easily fit situations in which vaccine-derived immunity is transferred either directly (as in the case of maternal antibodies against tetanus) or indirectly (as in the case of secondary spread of oral polio vaccines) between members of a population, or in which vaccines impart different levels of protection against infection, disease, or transmission (as in diphtheria, pertussis, and perhaps malaria).

The paper of Fox et al. (15) is also of importance because of its method and the nature of the conclusions which were dictated by that approach. Sufficient years have now elapsed for both the method and the conclusions to be reviewed in perspective.

Interest in applying the "magic" of herd immunity in disease control has encouraged mathematical research exploring the theoretical implications of the subject (6–8, 17–37). Though much of this work has been published in journals and in language unfamiliar to the medical and public health communities, its isolation has been reduced in recent years largely through the publications of Anderson and May and their colleagues (8, 17, 20, 21, 23, 24, 28, 29, 31, 33, 36).

It is the intent of this review to bring together the literature on the history, theory, and practical experience of herd immunity, to consider the variety of issues raised by the application of the concept to different diseases, and to consider how well current theory and practice correspond with each another.

HISTORY

The first published use of the term "herd immunity" appears to have been in a paper published in 1923 by Topley and Wilson titled "The spread of bacterial infection: the problem of herd immunity" (38). This was one of a classic series of studies by these authors on epidemics of various infections in closely monitored populations of laboratory mice (39). Topley and Wilson introduced the term in the following manner: "Consideration of the results obtained during the past five years . . . led us to believe that the question of immunity as an attribute of a herd should be studied as a separate problem, closely related to, but in many ways distinct from, the problem of the immunity of an individual host" (38, p. 243). After describing experiments showing that immunized mice had lower mortality rates from, and were less likely to transmit, *Bacillus enteritidis*, the authors concluded by posing an ". . . obvious problem to be solved. . . . Assuming a given total quantity of resistance against a specific bacterial parasite to be available among a considerable population, in what way should that resistance be distributed among the individuals at risk, so as best to ensure against the spread of the disease, of which the parasite is the causal agent?" (38, pp. 248–9). Wilson later recalled that he had first heard the phrase "herd immunity" in the course of a conversation with Major Greenwood (G. S. Wilson, London School of Hygiene and Tropical Medicine, personal communication, 1981); and Greenwood employed it in his 1936 textbook *Epidemics and Crowd Diseases* (40). Although these authors did not distinguish clearly between direct and indirect protection stemming from vaccine-derived immunity, later authors picked up the phrase and applied it in particular to the indirect protection afforded to nonimmune individuals by the presence and proximity of others who are immune.

That the presence of immune individuals could provide indirect protection to others was itself recognized at least as far back as the 19th century. Farr had noted in 1840 that "The smallpox would be disturbed, and sometimes arrested, by vaccination, which protected a part of the population . . ." (41). Such observations, that epidemics often came to an end prior to the involvement of all susceptibles, led in turn to a major epidemiologic controversy in the early years of this century. This controversy was between those who believed that epidemics termi-

nated because of changes in the properties of the infectious agent (e.g., loss of "virulence" resulting from serial passage) (42) and those who argued that it reflected the dynamics of the interaction between susceptible, infected, and immune segments of the population (43). Each argument was supported by observations and by mathematical reasoning (44). It was the latter explanation that won the day; and its simple mathematical formulation, the "mass action principle," which has become a cornerstone of epidemiologic theory, provides one of the simplest logical arguments for indirect protection by herd immunity.

The concept of herd immunity is often invoked in the context of discussions of disease eradication programs based on vaccination. It is significant that both Jenner (45) and Pasteur (46), key figures in the early development of vaccines, recognized the potential of vaccines to eradicate specific diseases, but neither appears to have considered the practical issues closely enough to have touched on herd effects. Furthermore, the major focus of eradication thinking in the first half of this century did not involve vaccines or vaccine-preventable diseases at all, but concerned vector-borne diseases, malaria in particular. This stemmed from the writings of Ross (47) who, in work on the dynamics of malaria, had deduced that it was not necessary to eliminate mosquitoes totally in order to eradicate the disease. Ross's so-called "mosquito theorem" was the first recognition of a quantitative threshold which could serve as a target for a disease elimination program. So powerful was the argument, and so influential was the tradition of quantitative thinking which it engendered, that the World Health Organization attempted global eradication of malaria before that of any other disease (48).[2] This tradition of mathematical epidemiology relating to vector-borne diseases has been repeatedly a source of important insights for the field of vaccination and herd immunity.

[2]The 1955 World Health Assembly recommended that the World Health Organization take the initiative in "a programme having as its ultimate objective the worldwide eradication of malaria." It was not until 1965 that the Assembly first declared "the worldwide eradication of smallpox to be one of the major objectives of the organization" (49).

THEORY

Three separate theoretical perspectives have been used to derive measures of herd immunity. Over recent years, these perspectives have converged into a general theory.

The mass-action principle

The theoretical basis of herd immunity was introduced by Hamer (43) in 1906 in the context of a discussion of the dynamics of measles. Hamer argued that the number of transmissions (he called it the "ability to infect") per measles case was a function of the number of susceptibles in the population. We can paraphrase his argument as:

$$C_{t+1}/C_t \text{ varies with } S_t, \qquad (1)$$

where S_t and C_t are numbers of susceptibles and cases, respectively, in some time period t, C_{t+1} is the number of cases in the succeeding time period, and C_{t+1}/C_t is, thus, the number of successful transmissions per current case (see figure 1). The time period used in this formulation is the average interval between successive cases in a chain of transmission, sometimes called the "serial interval" (50), which is approximately 2 weeks for infections such as measles and pertussis (see table 1). This relation can be expressed:

$$C_{t+1} = C_t S_t r, \qquad (2)$$

where r is a transmission parameter, or "contact rate," in effect the proportion of all possible contacts between susceptible and infectious individuals which lead to new infections. In order to simulate successive changes over time, the number of susceptibles is recalculated for each new time period as

$$S_{t+1} = S_t - C_{t+1} + B_t, \qquad (3)$$

where S_{t+1} is the number of susceptibles in

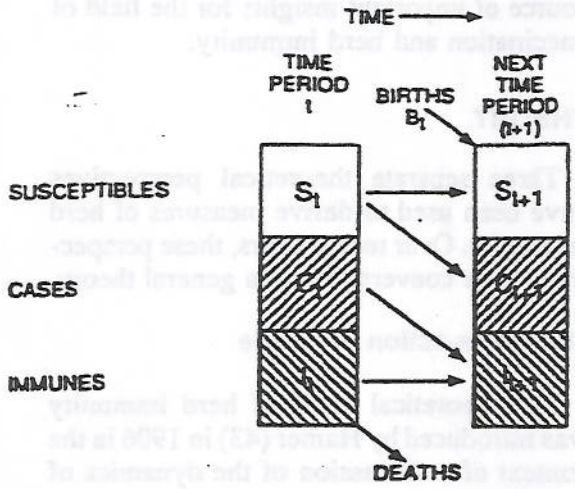

FIGURE 1. Relation between susceptibles (S), infectious cases (C), and immunes (I) in successive time intervals (t, $t+1$) in the simple discrete time mass action or Reed-Frost models. In each time period some (C_{t+1}) susceptibles become cases and the others remain susceptible. Each case is assumed to remain infectious for no more than a single time period (= serial interval). B_t individuals may enter as susceptible births during each time period (e.g., equation 3). Note that neither the simple mass action (equations 2 and 3) nor Reed-Frost (equation 9) equations include an explicit term for immunes. By implication, deaths prior to infection are not considered in these simplest models and the total population is assumed constant (i.e., in each period the same number of immunes die as susceptibles are born into the population).

the next time period and B_t is the number of new susceptibles added (e.g., born into) to the population per time period.

The relation in equation 2, that future incidence is a function of the product of current prevalence times current number susceptible, has become known as the epidemiologic "law of mass action" by analogy with the physical chemical principle that the rate or velocity of a chemical reaction is a function of the product of the initial concentrations of the reagents.[3] Often expressed as a differential (continuous time) rather than a difference (discrete time) equation, as here, this relation underlies most

[3]This analogy was apparently first made by Soper (51). The inspiration from physical chemistry is of more than passing interest in that it reflects a tradition among biomedical theorists to strive for the simplicity and elegance of the physical sciences. Not only mass action, but also the concepts of catalysis and of critical mass have close analogies in the behavior of infections, as mentioned below.

TABLE 1. Approximate serial intervals, basic reproduction rates (in developed countries) and implied crude herd immunity thresholds (H, calculated as $1 - 1/R_0$) for common potentially-vaccine-preventable diseases. Data from Anderson and May (8), Mcdonald (54), and Benenson (135). It must be emphasized that the values given in this table are approximate, and do not properly reflect the tremendous range and diversity between populations. They nonetheless give an appreciation of order-of-magnitude comparability

Infection	Serial interval (range)	R_0*	H^* (%)
Diphtheria†	2–≥30 days	6–7	85
Influenza‡	1–10 days	?	?
Malaria§	≥20 days	5–100	80–99
Measles‖	7–16 days	12–18	83–94
Mumps	8–32 days	4–7	75–86
Pertussis¶	5–35 days	12–17	92–94
Polio#	2–45 days	5–7	80–86
Rubella	7–28 days	6–7	83–85
Smallpox	9–45 days	5–7	80–85
Tetanus	NA°	NA	NA
Tuberculosis°°	Months–years	?	?

* R_0, basic case reproduction rate; H, herd immunity threshold defined as the minimum proportion to be immunized in a population for elimination of infection; NA, not applicable.

† Long-term infectious carriers of *Corynebacterium diphtheriae* occur. See the text for a discussion of the definition of immunity.

‡ R_0 of influenza viruses probably varies greatly between subtypes.

§ All these variables differ also between *Plasmodium* species. The serial interval may extend to several years. See the text for a discussion of implications of genetic subtypes.

‖ See the text for a discussion and variation in estimates of R_0 in table 5.

¶ See the text for a discussion relating to the definition of immunity in pertussis.

Distinct properties of different polio vaccines need to be considered in interpreting the herd immunity thresholds.

°° R_0 has been declining in developed countries; protective immunity is not well defined.

theoretical work on the dynamics of infections in populations (23, 52).

Figure 2 illustrates what happens when equations 2 and 3 are iterated and serves to illustrate several fundamental principles of the epidemiology of those acute immunizing infections (such as measles, mumps, rubella, chickenpox, poliomyelitis, pertussis, etc.) which affect a high proportion of individuals in unvaccinated communities.

First, the model predicts cycles of infection incidence, such as are well recognized for many of the ubiquitous childhood infections (figure 3). The incidence of infection cycles above and below the "birth" rate, or rate of influx of new susceptibles.

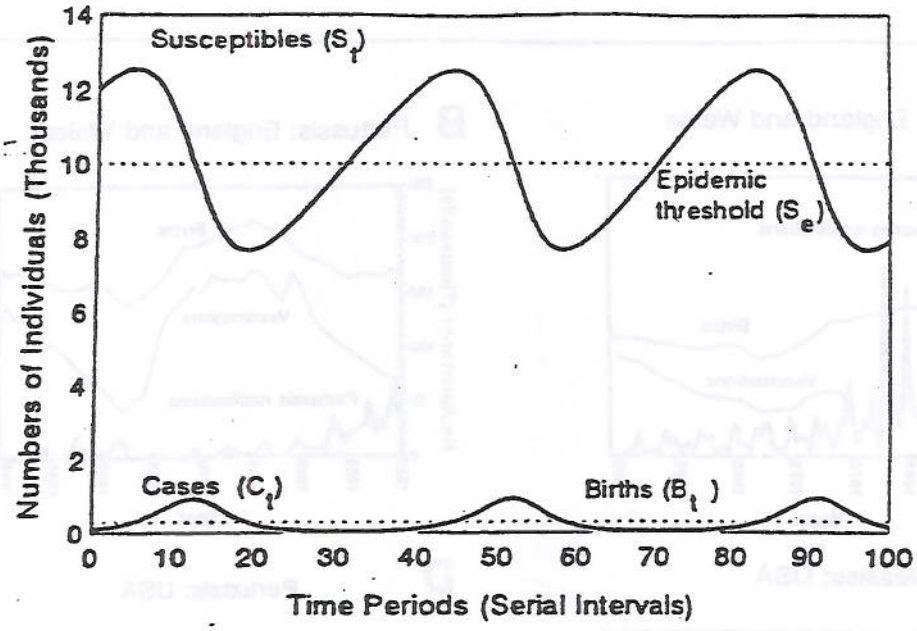

FIGURE 2. Mass action model. Results obtained on reiteration of equations 2 and 3. The illustrated simulation was based on 12,000 susceptibles and 100 cases at the start, $r = 0.0001$ and 300 births per time period. Note that the incidence of cases cycles around the birth rate and that the number of susceptibles cycles around the epidemic threshold: $S_e = 1/r = 10{,}000$.

Second, the number of susceptibles also cycles, but around a number which is sometimes described as the "epidemic threshold," S_e. Simple rearrangement of equation 1 to $C_{t+1}/C_t = S_t r$ reveals that this threshold is numerically equivalent to the reciprocal of the transmission parameter r; as incidence increases (i.e., $C_{t+1} > C_t$) when, and only when, $S_t > 1/r$; and, thus, $S_e = 1/r$. This important relation is implicit in Hamer's original paper (43), and was formalized as a "threshold theorem" in 1927 by Kermack and McKendrick (53). The principle may be illustrated by analogy with the physical concept of a "critical mass"—the epidemic threshold represents a critical mass (density per some area) of susceptibles, which, if exceeded, will produce an explosive increase in incidence of an introduced infection. The correspondence between the case and susceptible lines in figure 2 illustrates this relation.

Hamer and his successors used this logic to explain several aspects of the dynamics of measles and other childhood infections, such as cyclical epidemics, the persistence of susceptibles at the end of an epidemic, and the relation between the interepidemic interval and the time required for the number of susceptibles to reach the epidemic threshold (23, 43, 51, 52). Though it was not emphasized explicitly by the earlier authors, who dealt in numbers or "density," rather than proportions, of susceptibles, the epidemic threshold provides a simple numerical measure of a herd immunity criterion. If the *proportion* immune is so high that the *number* of susceptibles is below the epidemic threshold, then incidence will decrease. We can express this algebraically as:

$$H = 1 - S_e/T = 1 - 1/rT \qquad (4)$$

where T is the total population size, S_e is the epidemic threshold number of susceptibles for the population, and H is the herd immunity threshold, i.e., the *proportion* of immunes which must be exceeded if incidence is to decrease.

Figure 4 presents another way of illustrating the herd immunity threshold, i.e., in terms of the relation between the proportion immunized at birth and the ratio of the cumulative incidence during the postvaccination period to that during the prevaccination

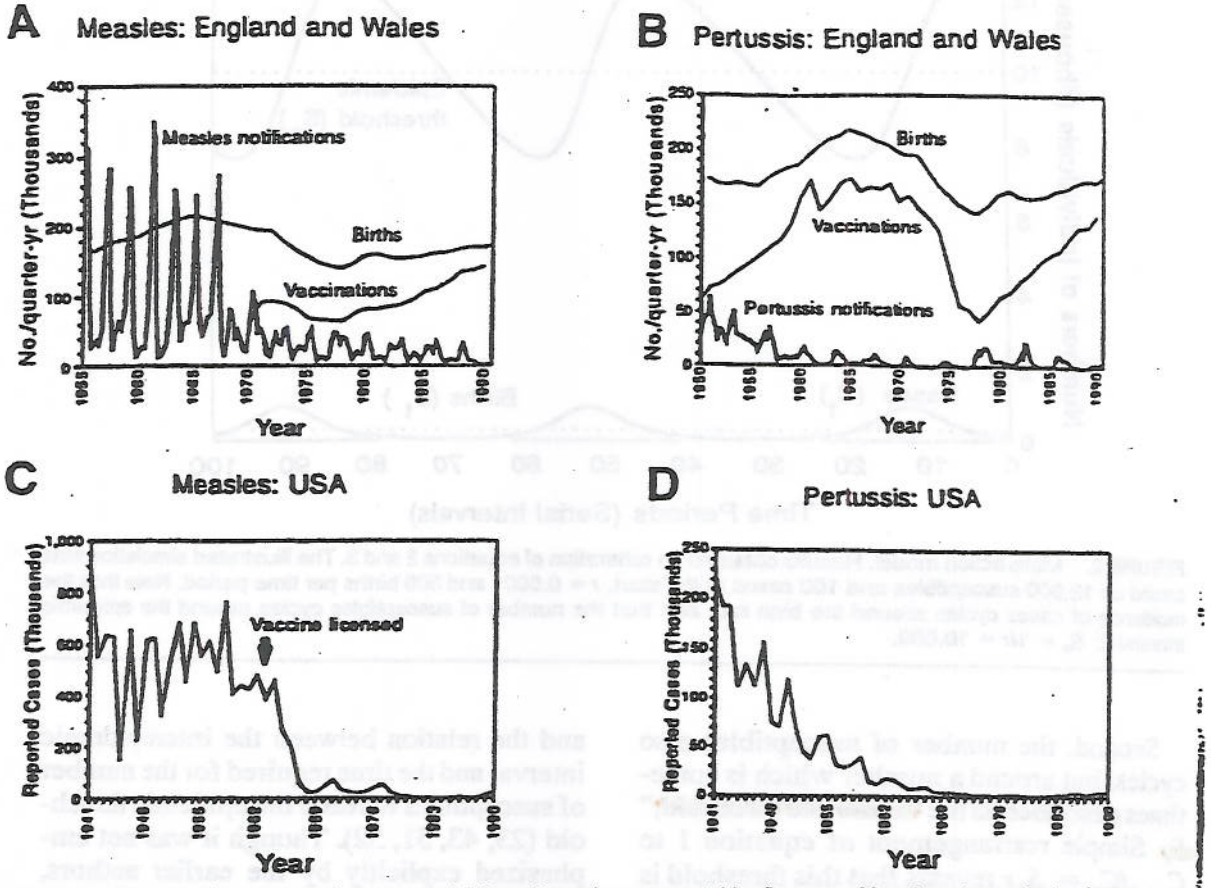

FIGURE 3. Reported incidence of common childhood vaccine-preventable diseases. Measles showed a tendency to biennial epidemics in England and Wales prior to vaccination (A). This pattern was less dramatic in data for the entire United States (C) because of the size and heterogeneity of the population (not all areas were in phase with one another). All areas showed a strong seasonal oscillation in addition to the biennial cycle. Pertussis shows a 3–4 year cycle with little obvious seasonality in the United Kingdom (B). This cycling is also seen in national data for the United States prior to 1970 (D). Notification efficiency was approximately 60% for measles in England and Wales prior to vaccination (55) but was considerably lower for pertussis and for both diseases in the United States.

period, either among those not immunized at birth (figure 4A), or in the entire population (figure 4B). Insofar as the immunization of individuals removes both susceptibles and potential sources of infection from the community, it will lead to a reduction in incidence rates and, hence, in cumulative incidence. If the proportion immunized at birth is maintained at or above the threshold, H, then the cumulative incidence is reduced to zero, indicating that the infection has been eliminated from the population.

It was only many years after Hamer that the wide use of vaccines meant that these epidemic and herd immunity thresholds could be considered as targets for intervention. If appropriate vaccination could prevent the number of susceptibles from reaching the epidemic threshold, then incidence should continue to decline, ultimately to extinction. Hamer's original principle implied the simplistic assumption of an homogeneous, randomly mixing population, like that of molecules in the ideal gasses for which the mass action principle was most appropriate. However, given the power of the analogy, elaboration of the theory was only a matter of time.

Case reproduction rates

If an infection is to persist, each infected individual must, on average, transmit that infection to at least one other individual. If

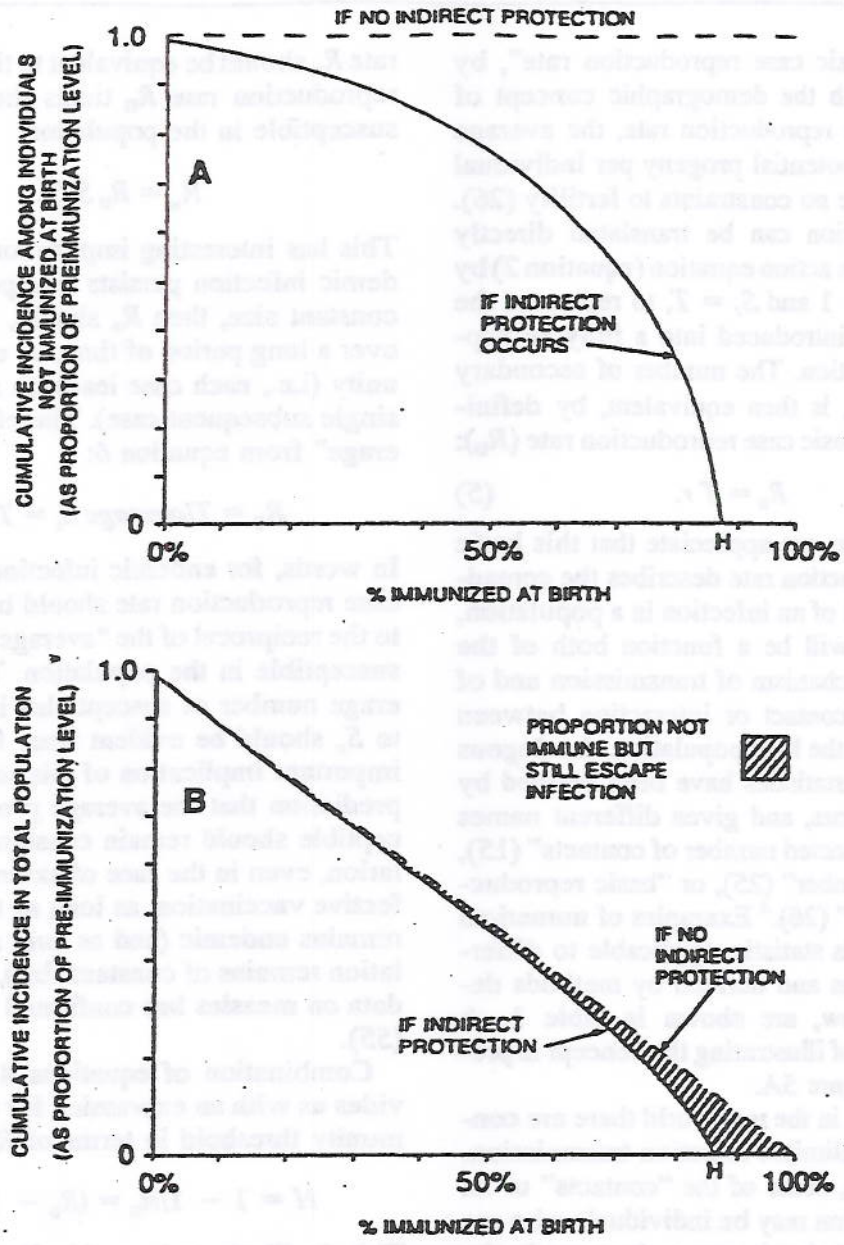

FIGURE 4. Cumulative incidence (e.g., per lifetime) of infection after a vaccination program as a proportion of prior cumulative incidence among individuals not immunized by the vaccine (A) and among the total population (B). In each diagram the *dotted line* refers to an infection for which the vaccine offers no indirect protection (e.g., tetanus vaccination of males) and the *solid line* refers to an infection for which the vaccine does impart indirect protection (e.g., measles). The vertical distance between the two lines reflects the nonimmunized individuals who escape infection as a proportion of all nonimmunized individuals (A) or of the total population (B).

this does not occur, the infection will disappear progressively from the population. This average number of actual infection transmissions per case is an extremely powerful concept, and has thus been discussed by many researchers. The fundamental statistic is one which was formulated originally by Macdonald (54), in the context of malaria studies, as the average number of secondary cases who contract an infection from a single primary case introduced into a totally susceptible population. He called this num-

ber the "basic case reproduction rate", by analogy with the demographic concept of the intrinsic reproduction rate, the average number of potential progeny per individual if there were no constraints to fertility (26). This definition can be translated directly into the mass action equation (equation 2) by letting $C_t = 1$ and $S_t = T$, to represent the single case introduced into a fully susceptible population. The number of secondary cases, C_{t+1}, is then equivalent, by definition, to the basic case reproduction rate (R_0):

$$R_0 = Tr. \quad (5)$$

On reflection, we appreciate that this basic case reproduction rate describes the spreading potential of an infection in a population, and that it will be a function both of the biologic mechanism of transmission and of the rate of contact or interaction between members of the host population. Analogous or identical statistics have been defined by several authors, and given different names such as "expected number of contacts" (15), "contact number" (25), or "basic reproduction number" (26).[4] Examples of numerical values of this statistic, applicable to different infections and derived by methods described below, are shown in table 1. A simple way of illustrating the concept is presented in figure 5A.

Of course, in the real world there are constraints to unlimited infection transmission. For example, some of the "contacts" of an infected person may be individuals who are already infected or immune. As a result, the average number of *actual* infection transmissions per case, in a real population, will be *less* than the basic case reproduction rate, and has been defined, again first by Macdonald (54), as the "net reproduction rate" R_n. Other authors have called this the "actual" or "effective" reproduction rate (23). This is illustrated in figure 5B. It is clear from figure 5 that the net reproduction rate R_n should be equivalent to the basic case reproduction rate R_0 times the proportion susceptible in the population:

$$R_n = R_0 S_t/T. \quad (6)$$

This has interesting implications. If an endemic infection persists in a population of constant size, then R_n should, *on average*, over a long period of time, be equivalent to unity (i.e., each case leads on average to a single subsequent case). Therefore, "on average" from equation 6:

$$R_0 = T/\text{average } S_t = T/S_e. \quad (7)$$

In words, for endemic infections, the basic case reproduction rate should be equivalent to the reciprocal of the "average" proportion susceptible in the population. That the average number of susceptibles is equivalent to S_e should be evident from figure 2. An important implication of this relation is the prediction that the average proportion susceptible should remain constant in a population, even in the face of extensive and effective vaccination, as long as the infection remains endemic (and as long as the population remains of constant size). Analysis of data on measles has confirmed this relation (55).

Combination of equations 4 and 7 provides us with an expression for the herd immunity threshold in terms of R_0:

$$H = 1 - 1/R_0 = (R_0 - 1)/R_0. \quad (8)$$

This is illustrated graphically in figure 6 which shows the implications for persistence or eradication of infections depending on the proportion of immunes in the population.[5]

The Reed-Frost heterogeneous population simulation approach

The paper by Fox et al. (15) cited in the introduction has been one of the most fre-

[4] Different symbols have been used for the statistic by different authors. The original work by Macdonald (54) employed Z_0 for the basic reproduction rate. Several authors have noted that the statistic is not a proper rate, but that term is now imbedded in the literature (26).

[5] This important relation was published explicitly first by Dietz (18), in 1975, though it is implicit in some earlier work, in particular a graph published by Smith (56) in 1970.

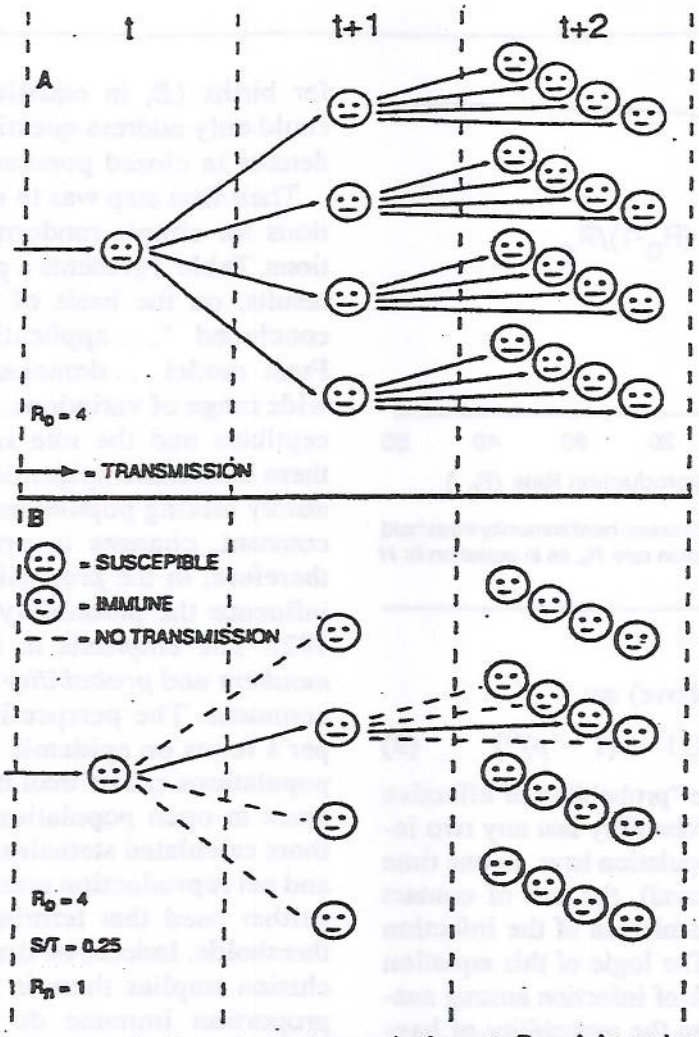

FIGURE 5. Cartoon illustrating implications of a basic reproduction rate $R_0 = 4$. In each successive time (serial) interval, each individual has effective contact with four other individuals. If the population is entirely susceptible (A), incidence increases exponentially, fourfold each generation (until the accumulation of immunes slows the process). If 75% of the population is immune (B), then only $S/T = 25\%$ of the contacts lead to successful transmissions, and the net reproductive rate $R_n = R_0 (S/T) = 1$.

quently cited references on herd immunity. This paper is of historical interest, and also of interest because of its theoretical argument and conclusions.

The appearance of the Fox et al. paper in 1971 was significant. Four years before, in 1967, the World Health Organization had declared its intention of eradicating smallpox from the world within 10 years, and the United States Public Health Service had declared its intention of eradicating measles from the United States within 1 year (57). Both of these tasks were to be achieved by the induction of herd immunity with vaccines. By 1971, the initial successes and failures of these programs were on record (e.g., figure 3C), and Fox et al. set out to explain them.

They based their theoretical argument not on the mass action arguments outlined above, but on an alternative approach, rooted in the Johns Hopkins University School of Hygiene and Public Health (58). This model, named the Reed-Frost for its developers Lowell Reed and Wade Hampton Frost, assumes the same discrete time schema illustrated in figure 1 but proposes an alternative to the mass action equa-

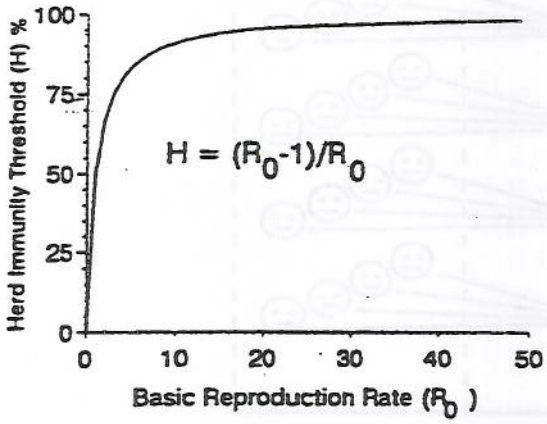

FIGURE 6. Relation between herd immunity threshold (H) and basic reproduction rate R_0, as in equation 8: $H = 1 - 1/R_0$.

tion (equation 2 above) as:

$$C_{t+1} = S_t\{1 - (1 - p)^{C_t}\} \qquad (9)$$

where p equals the "probability of effective contact," or the probability that any two individuals in the population have, in one time period (serial interval), the sort of contact necessary for transmission of the infection in question (58). The logic of this equation is such that the risk of infection among susceptibles is equal to the probability of having effective contact with at least one infectious case.[6] This model had traditionally been applied to simulate epidemics in closed populations (with no births or influx of susceptibles). Fox et al. continued this tradition, and thus calculated susceptibles for successive time periods as

$$S_{t+1} = S_t - C_{t+1}. \qquad (10)$$

This is important, as, by omitting any term for births (B_t in equation 3), the authors could only address questions relating to epidemics in closed populations.

Their first step was to explore these equations for simple randomly mixing populations. Table 2 presents a portion of the initial results, on the basis of which the authors concluded "... application of the Reed-Frost model ... demonstrates that, over a wide range of variations, the number of susceptibles and the rate of contact between them determine epidemic potentials in randomly mixing populations. If these are held constant, changes in population size and, therefore, in the proportion immune do not influence the probability of spread" (15, p. 182). The emphasis in this conclusion on *numbers* and *probability of spread* deserves comment. The perspective reflects the paper's focus on epidemic potential in closed populations rather than on infection persistence in open populations. Though the authors calculated statistics analogous to basic and net reproduction rates (see table 2), they neither used that terminology nor derived thresholds. Indeed, on the surface, their conclusion implies there is no threshold ("the proportion immune do not influence the probability of spread"), though this is a consequence of the assumption that "numbers of susceptibles and the rate of contact" are held constant. But, given the definition of the Reed-Frost contact rate as the probability that *any* two individuals have effective contact in one time period, it is unreasonable to consider alteration of population size without accepting its implications for some consequent change in contact probabilities. (For example, the probability for any two people *chosen at random* in a small community to meet, by chance, in 1 week, may be 0.1, but this probability will surely be smaller if they live in a very large population). Viewed from this perspective, the authors' first conclusion, as quoted above, appears almost spurious.

The paper then took a crucially important step. The authors explored an alternative to the basic assumption of homogeneous random mixing, which had been implicit in all

[6] If the same value is substituted for r in equation 2 and p in equation 9, the mass action predicts a higher number of successive cases than does the Reed-Frost for any given S_t and C_t. This is because the mass action equation does not correct for the fact that multiple infections on a single susceptible can lead to only a single subsequent case. It can be shown by the binomial expansion that the Reed-Frost model approximates the mass action if p is small, in which case the Reed-Frost p and the mass action r become the same statistic (59). This is reasonable in that as p is reduced, the probability of a susceptible contacting more than one case per serial interval (e.g., p^2 is the probability of contacting two cases, etc.) becomes vanishingly small.

TABLE 2. Extract from a table published by Fox et al. (15) to illustrate the behavior of infections in a randomly mixing population, as predicted by the Reed-Frost model

Initial population composition				"Probability of effective contact"	Expected number of effective contacts by case in first interval		Probability of no spread
Susceptibles (S)	Cases (C)	Immune S (I)	Total (N)	(p)	With susceptibles pS*	Total p(N − 1)†	(1 − p)^10‡
10	1	0	11	0.2	2	2	0.11
10	1	5	16	0.2	2	3	0.11
10	1	5	16	0.133	1.3	2	0.23

* Analogous to the net reproduction rate, R_n.
† Analogous to the basic reproduction rate, R_0.
‡ The probability that all 10 susceptibles fail to have contact with the single index case.

modeling arguments to that time. They set up a structured community in which 1,000 individuals were separately assigned family, school, and social groupings, each of which had a different internal contact probability. By using Monte Carlo techniques, they simulated the consequences of introducing infections into such populations with and without opportunities for special mixing within and between the social groups. Table 3 presents a portion of the results of these simulations, which led the authors to conclude: "Free living populations of communities are made up of multiple and interlocking mixing groups, defined in such terms as families, family clusters, neighborhoods, playgroups, schools, places of work, ethnic and socioeconomic subgroups. These mixing groups are characterized by different contact rates and by differing numbers of susceptibles. The optimum immunization program is one which will reduce the supply of susceptibles in all subgroups. No matter how large the proportion of immunes in the total population, if some pockets of the community, such as low economic neighborhoods, contain a large enough number of susceptibles among whom contacts are frequent, the epidemic potential in these neighborhoods will remain high. Success of a systematic immunization program requires knowledge of the age and subgroup distribution of the susceptibles and maximum effort to reduce their concentration throughout the community, rather than aiming to reach any specified overall proportion of the population" (15, p. 186). While the argument that social structure is important in determining patterns of infection is compelling, two points in this con-

TABLE 3. Relative frequency distributions of epidemic sizes predicted by the Reed-Frost model, assuming different structures to a population of 1,000 persons. Data are based on 100 stochastic simulations under each set of conditions, as published by Fox et al. (15)

Mixing groups	Within group contact (p value)	Total number of cases per epidemic (%)										Mean epidemic size
		1	2	3	4	5–9	10–19	20–29	30–39	40–59	60–79	
Total community	0.002	82*	15	2	1							1.2†
Total community	0.002	22	18	34	8	17	1					3.3
Families, [62]‡	0.5											
Total community	0.002											
Families, [62]	0.5	11	6	26	23	23	9	1	1			5.6
Playgroups [24]	0.1											
Total community	0.002											
Families, [62]	0.5											
Playgroups [24]	0.1	23	4							28	45	45.0
Nursery school	0.1											

* Thus, 82 of the 100 epidemics simulated under these conditions (in this case a randomly mixing community with probability of effective contact, $p = 0.002$), terminated after a single case.
† The average total number of cases in all 100 simulated epidemics was 1.2.
‡ The numbers in brackets reflect the numbers of families, playgroups, and nursery schools in the simulated populations.

clusion are less clear. First, the statement that it is important to reduce the supply of susceptibles in *all* subgroups is not strictly supported in the paper's theoretical results; indeed, it is intuitively reasonable, and was later demonstrated in theory (see below), that targeting vaccination to groups with high contact probabilities can be more efficient (in the sense of minimizing the total number of vaccinations required) in reducing disease than is uniform coverage of an entire population. Second, the emphasis on curbing epidemic spread remains. Although Fox et al. considered their approach "... relevant to programs of systematic immunization ... which have as their ultimate goal elimination of the causative agent from the country" (15, p. 186), it was most relevant to epidemics in closed populations, as it had no provision for examining the implications of a constant influx of susceptibles into the population, as by birth.

The Fox et al. paper deserves its considerable influence. Its break from the tradition of random mixing populations was a crucially important development. Its theory was born of practical experience and disappointment with progress in measles control in the United States, and its tone was pessimistic and practical, compared with most of the past (and subsequent) literature on herd immunity, which has trended to emphasize simple thresholds. As we shall see, the paper still proves to be wise counsel.

Recent theoretical developments

The credibility of the simple formulations of herd immunity thresholds is weakened by the fact that the logic and formulae are based on obviously simplistic assumptions. In particular, the basic mass action models assumed that populations are homogeneous, with no differences by age, social group, or season, and that they mix at random. Mathematically inclined workers have taken these failings as a challenge to adapt the theory to more realistic assumptions.

The estimation of R_0. The centerpiece of research on herd immunity has been the linking of the mass action and basic case reproduction rate theories. The crucial insight appeared in a 1975 paper by Dietz (18) which demonstrated that, if one assumes a stable population in which the mortality rates and the incidence rates of infection are both independent of age, then

$$R_0 = T/S_e = 1 + L/A. \qquad (11)$$

where L is defined as the average expectation of life and A is the average age at infection.[7] Mathematical proofs of this relation have been presented by several authors (18, 23, 25, 27). The derivations assume an exponential distribution of the population by age and age-independent incidence rates of infection (figure 7A).[8] The relation can take an even simpler form if the population is assumed to have a rectangular age distribution (figure 7B), in which case

$$R_0 = L/A. \qquad (12)$$

This latter relation can be illustrated neatly if we recall that R_0 is equivalent to the reciprocal of the proportion susceptible at equilibrium ($R_0 = T/S_e = 1/s_e$), and assume that everyone is infected at exactly age A, the average age at infection, and dies at exactly age L, the average expectation of life (figure 7B). Assuming this rectangular age structure, the proportion susceptible is A/L; thus $R_0 = L/A$. On this basis, we might conclude that the higher crude estimates of R_0 implicit in equation 11 should in general be more appropriate for developing countries, with pyramidal or exponential age distributions (figures 7A and C), and the lower estimates of equation 12 for developed countries (figures 7B and D).

[7] This insight represents another contribution stemming from the traditions of the mathematics of vector-borne diseases (Dietz's paper (18) was on arthropod-borne viruses) and of physical chemistry (the assumption of an age-independent incidence rate is the basis of the so-called "catalytic models" (60)).

[8] In brief, if μ is the death rate and λ is the force (person-time incidence rate) of infection, then the average duration of life is $1/\mu = L$ and the average duration of susceptible life is $1/(\lambda + \mu)$. As $R_0 = 1/(\textit{proportion susceptible})$, $R_0 = (\lambda + \mu)/\mu = 1 + \lambda/\mu$. If μ is small compared to λ, then this expression is close to $1 + L/A$.

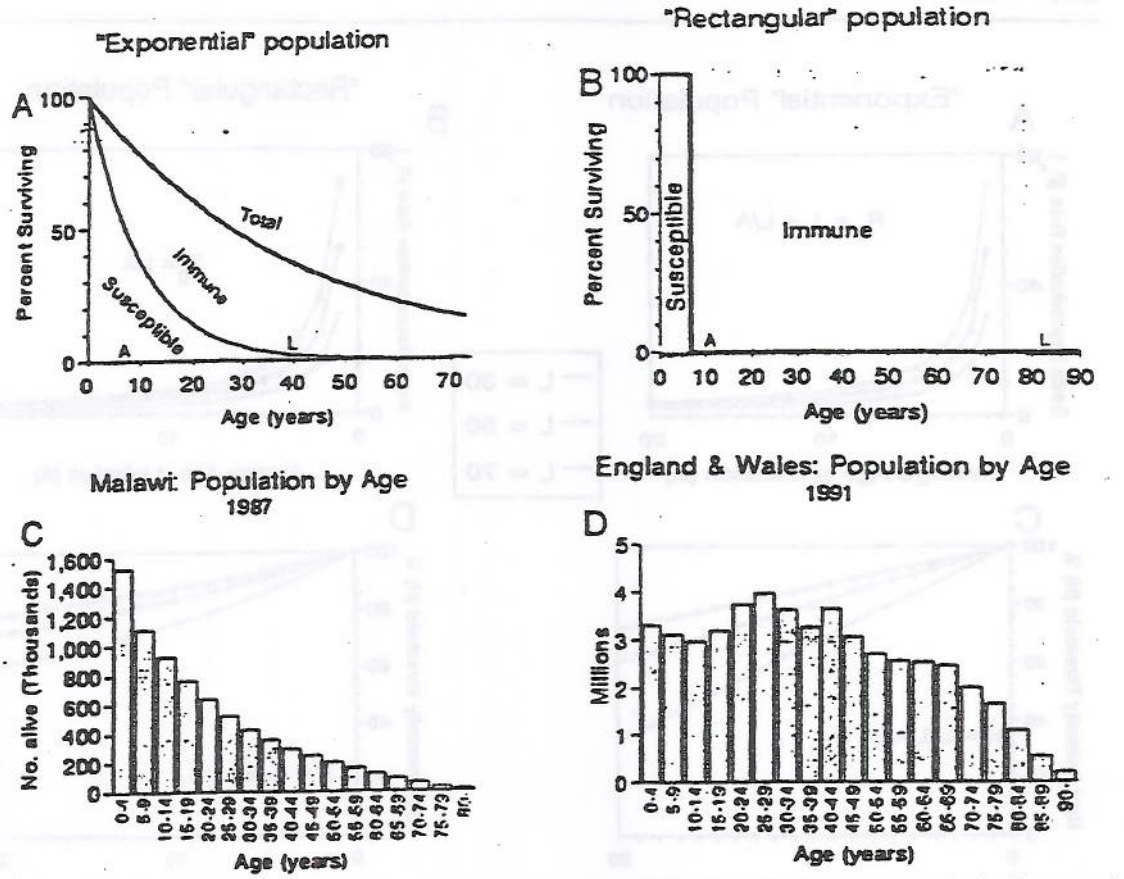

FIGURE 7. Schematic diagrams of exponential (A) or rectangular (B) age distributions compared with current population distributions in Malawi (C) and England and Wales (D). The exponential model (A) assumes infection and constant death rates at all ages. The average age at infection and average expectation of life are *A* and *L* years, respectively. In the rectangular model, all individuals are assumed to become infected at age *A* and to die at age *L*.

Equations 11 and 12 may be combined with the basic herd immunity expression (equation 8) to give relations between crude basic reproduction rates, herd immunity thresholds, and average age at infection, as shown in figures 8A–8D. The availability of such expressions has made it a straightforward matter to estimate crude basic reproduction rates and herd immunity thresholds for a variety of diseases of childhood (see table 1). Beyond that, they have opened the way to explorations of more realistic (and complicated) sets of assumptions.

Age-related effects. The simple mass action and Reed-Frost models make no provision for the fact that individuals pass through periods of different infection risk as they age. The inclusion of this factor requires compartmentalization of the population by age groups as well as by infection status (i.e., with maternal immunity, or susceptible, or latent, or infectious, or with active immunity). Assumptions must then be made as to how the risk of infection, within each age group in each time period, is a function of the prevalence of infectious cases in the same and other age groups at that time. A general scheme for this approach is presented in figure 9. Several investigators have tackled the problem and have thus been able explore the effects of different age-specific contact patterns, and vaccination strategies, within simulated populations (7, 19, 23, 36). Not surprisingly, the simple elegance of the basic mass action model has been lost, and the results have

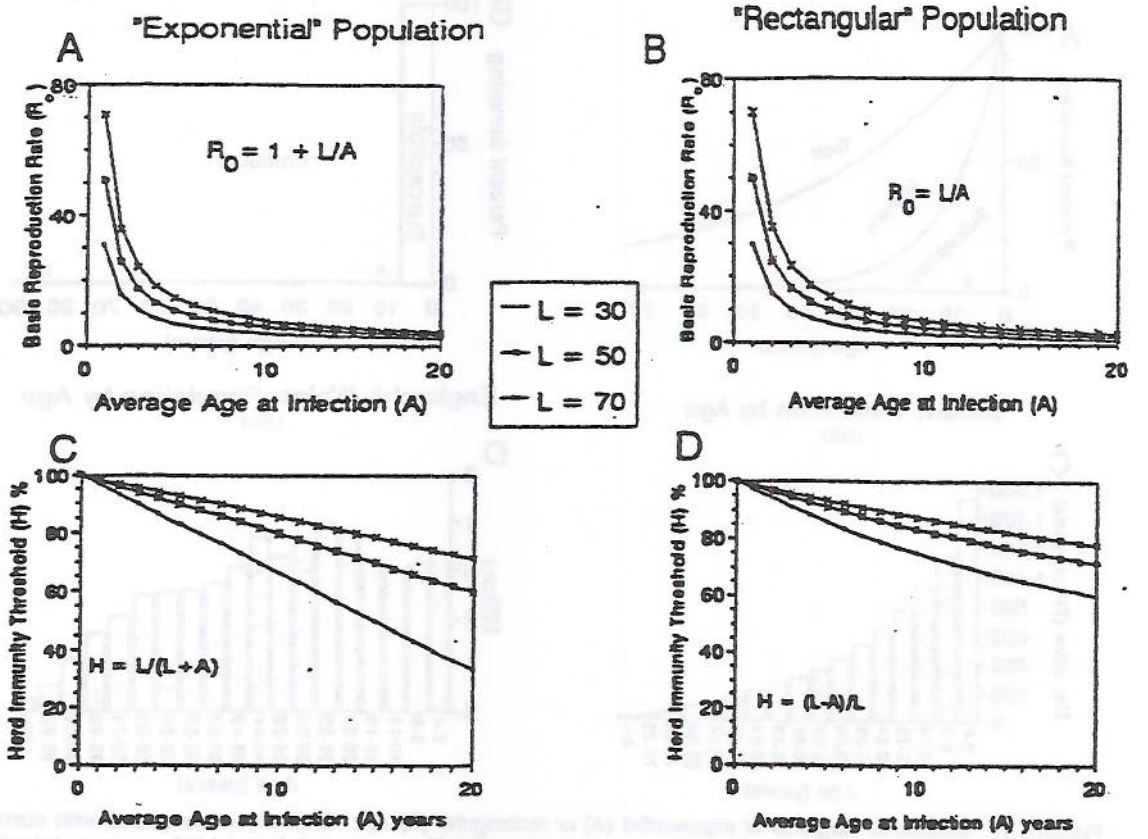

FIGURE 8. Relation between R_0 (basic case reproduction rate), H (herd immunity threshold), A (average age at infection), and L (average expectation of life), based on exponential (A and B) or rectangular (C and D) age distribution assumptions, derived from equations 8, 11, and 12.

become more complex, and less easily generalized, as the number of variables has increased. On the other hand, several principles have emerged.

Inclusion of *maternal immunity* (transplacentally-acquired immunoglobulin G) in the models serves to increase slightly the estimates of basic reproduction rates and herd immunity thresholds calculated from equations 11 and 12 (23). This is intuitively reasonable in that, as far as an infectious agent is concerned, an individual does not really enter the population until he or she has lost maternal antibody protection (and, thus, the A and L parameters in equations 11 and 12 are, in effect, overestimates). The basic equations can thus be adapted to adjust ages as though they were calculated from the average age of losing maternal immunity, M (on the order of 0.5 years for measles but less for many other infections), rather than from birth, for example,

$$R_0 = 1 + (L - M)/(A - M). \quad (13)$$

Another use of this approach has been to explore the implications of *vaccinating at different ages*. Selection of the optimal age for vaccination is dependent on several factors, including the duration of interfering maternally-acquired antibodies, logistic requirements of the health services, and the need to protect children prior to exposure to risk. The issue is complicated further insofar as vaccination itself may reduce infection risks, and, hence, expand the "window" period prior to any given level of cumulative incidence. On the other hand, age at vaccination is related inversely to the reduction of susceptibles in the population, and, hence, affects estimates of herd immunity thresholds. This is easily described in terms of the