Epidemiol. Infect. (2002), **129**, 147–153. © 2002 Cambridge University Press
DOI: 10.1017/S0950268802007148 Printed in the United Kingdom

A clarification of transmission terms in host-microparasite models: numbers, densities and areas

M. BEGON[1]*, M. BENNETT[2], R. G. BOWERS[3], N. P. FRENCH[2], S. M. HAZEL[2] AND J. TURNER[2]

[1] *Centre for Comparative Infectious Diseases and Population and Evolutionary Biology Research Group, School of Biological Sciences, The University of Liverpool, Liverpool, UK*
[2] *Faculty of Veterinary Science, The University of Liverpool, Liverpool, UK*
[3] *Department of Mathematical Sciences, The University of Liverpool, Liverpool, UK*

(*Accepted 7 March 2002*)

SUMMARY

Transmission is the driving force in the dynamics of any infectious disease. A crucial element in understanding disease dynamics, therefore, is the 'transmission term' describing the rate at which susceptible hosts are 'converted' into infected hosts by their contact with infectious material. Recently, the conventional form of this term has been increasingly questioned, and new terminologies and conventions have been proposed. Here, therefore, we review the derivation of transmission terms, explain the basis of confusion, and provide clarification. The root of the problem has been a failure to include explicit consideration of the area occupied by a host population, alongside both the number of infectious hosts and their density within the population. We argue that the terms 'density-dependent transmission' and 'frequency-dependent transmission' remain valid and useful (though a 'fuller' transmission term for the former is identified), but that the terms 'mass action', 'true mass action' and 'pseudo mass action' are all unhelpful and should be dropped. Also, contrary to what has often been assumed, the distinction between homogeneous and heterogeneous mixing in a host population is orthogonal to the distinction between density- and frequency-dependent transmission modes.

INTRODUCTION

Much of what is understood about the dynamics of infectious disease has been derived from investigations of mathematical models of disease dynamics or the application of those models in medical, veterinary or ecological contexts [1]. The transmission of infection from infectious to susceptible hosts is arguably the driving force in the dynamics of any infectious disease. A crucial element of the mathematical models, therefore, and hence of an understanding of disease dynamics, has been the 'transmission term' describing the rate (per unit time) at which susceptible hosts are 'converted' into infected hosts by their contact with infectious material. Recently, the conventional form of this term, and conventional terminologies, have been increasingly questioned [2], and new terminologies, and even new conventions, have been proposed. No consensus has been arrived at, however, giving rise, potentially, to confusion and an unproductive lack of uniformity.

Here, we review the derivation of transmission terms, explain the basis of the confusion, and thus, hopefully, provide clarification. These transmission terms describe a process in which there is no explicit spatial behaviour (no 'movement parameters'), although spatial distribution and movement is implicitly

* Author for correspondence: School of Biological Sciences, Nicholson Building, The University of Liverpool, Liverpool L69 3GS, UK.

incorporated. Further, they refer to populations of hosts within which there are no distinct sub-classes in terms of the biological characteristics involved in transmission, although where there are such sub-classes, separate transmission terms of the type discussed here may be applied to transmission within each class and between classes.

A DERIVATION FROM NUMBERS OR DENSITIES?

De Jong and his collaborators [2, 3] have argued that at the heart of the disagreements surrounding transmission terms has been variation and confusion in whether numbers or densities of hosts are used in their derivation. They claim that densities should be used, and that those who have used numbers (e.g. [4–7]) have been wrong to do so. We disagree. First, in host-pathogen dynamics, numbers of hosts, densities of hosts (numbers per unit area) and the area occupied by the host population may all vary – over time and from population to population. If a formulation (e.g. [2, 8–10]) is based on densities alone, it is impossible to distinguish variations in numbers from variations in areas-occupied. On the other hand, a derivation can only be based on numbers alone if it is assumed at the outset that the area occupied by populations is constant, which is not true as a generalization. Moreover, given that area may vary, the 'balance equations', to which the transmission terms contribute, have to be constructed on the basis of counting individuals, which have a continuing integrity, rather than balancing densities, which do not. Hence, in the following section, we derive transmission terms using arguments in which both numbers *and* areas enter explicitly. Terms based on densities can then be derived from these.

A BASIC TRANSMISSION TERM

The argument is developed for directly transmitted infections, in which susceptible hosts become infected by direct 'contact' with infectious hosts, that is, without the intervention of a vector species or free-living infectious particles. The transmission term appears traditionally [4] both in equations describing the changing numbers of susceptible hosts, S, as a 'loss' term, and also in equations describing the changing numbers of infected hosts, I, as a 'gain' term, counteracted by the loss of infecteds through death and recovery. In what follows, for clarity, we deal only with the infected-host equation, and we omit the loss term: dI/dt will refer only to the rate of increase, through new infection, in the number of infecteds.

This rate increases with the number of susceptible hosts, S, 'available' to be converted into infected hosts, which must then be multiplied by a per capita rate, conventionally referred to as the 'force of infection'. The force of infection is the product of (i) the rate of contacts, c, which are of an appropriate type for transmission to be possible if one of the hosts is infectious, (ii) the probability, p, that a contact is indeed with an infectious host, and (iii) the probability, ν, that contact between an infectious and a susceptible host does in fact lead to transmission (i.e. is 'successful'). This gives rise to the following basic equation:

$$dI/dt = Scp\nu. \tag{1}$$

The probability of successful transmission, ν, is usually assumed to be constant for any given host-pathogen combination. The probability that the contact is with an infectious host, p, is usually assumed to be I/N, the prevalence of infection within the population, where N is the total number of hosts in the population. This clearly depends on the assumption that what applies globally, as a proportion, to the whole population also applies 'locally', as a probability, to given susceptibles within the population. Transmission terms (and forces of infection) are then usually distinguished on the basis of the rate of contact, c.

DENSITY- AND FREQUENCY-DEPENDENT TRANSMISSION

The first, and most frequently assumed possibility is that $c = \kappa N/A$; that is, the rate of contact increases directly with the density of the population, N/A (where A is the area occupied by the population) scaled by a constant, κ, which varies with the combination of host and pathogen. The product $\kappa\nu$ is usually referred to as β, the transmission coefficient. This leads to the following equation:

$$dI/dt = S\kappa\,(N/A)\,(I/N)\,\nu = \beta SI/A. \tag{2}$$

Thus, not only does the contact rate increase with the overall density of the host, N/A; the per capita force of infection also increases with the density of

infecteds, I/A. Equation (2), then, is often said to describe 'density-dependent transmission', a terminology to which we subscribe. (Note that, in contrast to previous derivations, a dependence on density can unambiguously be introduced into an argument based on numbers, precisely because the area occupied by the population is included explicitly.)

A commonly assumed alternative possibility is that $c = \eta$; that is, the rate of contact is constant irrespective of the density of the population. The product $\eta\nu$ may be referred to as β', another transmission coefficient (albeit with different dimensions to β – see below). This leads to the following equation:

$$dI/dt = S\eta\,(I/N)\,\nu = \beta' SI/N. \tag{3}$$

Here, the per capita force of infection increases with the prevalence of infection, I/N, which might also be called the 'frequency' of infecteds; and the contact rate, being constant, may be said to increase with the 'frequency' of contacts, though this is tantamount to saying that the contact rate increases with the contact rate. Be that as it may, equation (3) is often said to describe 'frequency-dependent transmission', a terminology to which we also subscribe (in part because of our wish not to introduce new terms – such as the logical 'prevalence-dependent transmission', or 'proportionate mixing' [7] – if this can be avoided).

Thus, density- and frequency-dependent transmission may be distinguished on the density- and frequency-dependences of their forces of infection, and also (stretching terminologies) on the density- and frequency-dependences of their contact rates. In the past, different authors have used different distinctions in defining the contrasting transmission types, or (more often) they have been less than explicit in justifying their use of the terms. We believe, nonetheless, that 'density- and frequency-dependent transmission' are sufficiently well tied to the fundamental distinctions between the two types for their use as a generally accepted terminology to be warranted, especially in view of the unacceptability of alternative terms (see below).

TRANSMISSION TERMS BASED ON DENSITIES?

Equations (2) and (3) can also readily be expressed in terms of densities (numbers per unit area) by replacing I with iA and so on (where the lower case refers to the density-equivalent of the number referred to by the capital letter) – but only provided that the area occupied by the population is assumed to remain constant over time. The alternative of allowing area to vary over time is examined briefly in a later section. For now, the density (constant area) equivalent of equation (2) is:

$$di/dt = \beta si, \tag{4}$$

while that for equation (3) is:

$$di/dt = \beta' si/n. \tag{5}$$

It follows from this that the numbers equations (2) and (3) may be used as they stand; but the density equations (4) and (5), although they have often been used (e.g. [2, 8–10]) should only strictly be used if an assumption of constant area over time is acknowledged.

MASS ACTION TRANSMISSION?

Density-dependent transmission has frequently been described as 'mass action transmission' [10], by analogy with the binary collision of gas particles in a perfect gas exhibiting Brownian motion and subject to the Law of Mass Action (apparently going back to Hamer [11]). As pointed out by De Jong et al. [2], however, the numbers form of this has generally been quoted not as equation (2) but as:

$$dI/dt = \beta^* SI, \tag{6}$$

where the * has been inserted to indicate, again, that the dimensions of β are different from those in equation (2). Specifically, $\beta^* = \beta/A$, which is therefore only constant if A is constant – not only *within* a given population over time, but also *between* different populations being compared at the same time. It is only under these (unlikely) circumstances that equation (6) can be used as a compact form of equation (2). Otherwise, equation (2) reflects the biological reality, absent from equation (6), that for given values of S and I, dI/dt will be greater in a population occupying a smaller area, in which hosts are more likely to make contact with one another. Similarly, equation (4) can only be described as mass action transmission if the population is assumed to occupy 'an arena of fixed size' [10]. Equation (2) may therefore lay claim to describing 'mass action transmission', though this claim appears never to have been made. But equations (4) and (6), which have frequently been said to

describe mass action transmission, only do so if constancy of A within (in the case of equation (4)) or both within and between populations (equation (6)) are assumed. We contend, therefore, that use of the term 'mass action' should be discontinued on the grounds of inevitable uncertainty as to whether it is being applied to equation (2), which would be correct but out of line with past practice, or equation (4) or (6), which would be in line with much past practice but either limited in its application or incorrect.

Interestingly, the literature of chemical kinetics suggests that the first statement of the Law of Mass Action was due to Guldberg and Waage [12] and took the following form. 'The velocity of a reaction, at a given temperature, is proportional to the product of the concentrations of the reacting substances.' Thus concentrations (which are equivalent to densities) are invoked. However later work [13] parallels points made here in recognizing that care must be exercised in dealing with the relation between numbers of molecules and concentrations in reactions at different or varying volume.

TRUE AND PSEUDO MASS ACTION?

In their series of papers in which they criticised the confusion that may arise from 'numbers' derivations of mass action transmission, De Jong and colleagues [2, 3] went on to argue that the phrase 'mass action' should actually be applied to the transmission term $\beta' SI/N$ rather than $\beta^* SI$. They did so because this was the 'numbers' transmission term that emerged from a particular density-dependent derivation of their own, based initially on densities. However, as part of their derivation, they suggested that each individual occupies a characteristic area, such that the area occupied by a population is directly proportional to the size of that population. In this case (and only in this case) A and N are interchangeable (with appropriate re-scaling), leading effectively to the conclusion that equation (2) implies equation (3) with $\beta = \beta'$. From this, De Jong et al. suggested that '$dI/dt = \beta' SI/N$' should be referred to as '*true* mass action transmission', whereas '$dI/dt = \beta^* SI$' should be referred to as '*pseudo* mass action transmission', a suggestion that received some initial support [14, 15].

Contrary to their argument, however, the area occupied by many populations does not increase as the numbers increase, and A and N are not therefore generally interchangeable. Hence, we contend that their argument is flawed and that adoption of the terms 'true' and 'pseudo mass action' is not justified. Indeed, referring to '$dI/dt = \beta' SI/N$' as 'true mass action transmission' is particularly misleading, since in general it derives not from a 'mass action' argument but from one with a constant contact rate (equation (3)). Equation (2) is only equivalent to this in the special case where N and A are interchangeable, because this means that density and thus contact rate *are* constant.

NUMBERS, DENSITIES AND AREAS OCCUPIED

Confusing numbers and densities in an argument, and ignoring area, may seem difficult to understand. Indeed, criticisms of this have come from a veterinary background (e.g. [2]), where it is natural to expect larger populations to occupy larger areas: larger populations of pigs occupy more pens, larger herds of cattle occupy more fields. However, the derivations that have been criticised have often concerned human and other populations which have been assumed to remain constant in size, at least on time-scales comparable to those over which disease dynamics are studied ('epidemiological' rather than 'ecological' studies [16]). In such cases, it is perhaps natural to also assume, implicitly, that the area occupied by populations is always the same. Likewise, ecological studies, in which host population size does vary, have focused on wildlife populations where investigations are usually carried out on an experimental or observational 'grid' or 'sampling unit' within a much larger host population [14, 17]. Thus, numbers and densities within a population *are* effectively the same, since the area of the sampling unit is constrained by practicalities never to vary in size.

In practice, therefore, the assumption of constant area within a population may have been a safe one in many epidemiological and wildlife studies. Even in these types of systems, however, there are effectively no grounds for the constant-area assumption when different populations are compared. There seems little doubt that equation (2) is the more general, and in that sense more correct, equation, which should be replaced by equation (6) only when the constant-area assumption is made explicitly.

In passing, note that it has also been claimed [6, 18] that in populations (e.g. humans) where numbers remain effectively constant and are not affected by

disease, density- and frequency-dependent models are equivalent; that is, equations (3) and (6) are the same because N is constant. Including area, however, makes it apparent that this, too, is true (i.e. equations (2) and (3) are the same) only if it is additionally assumed not only that the area occupied by any given population remains constant, but that all populations that might be compared occupy the same area.

Explicit inclusion of area, moreover, raises the question of what precisely is meant by 'the area occupied by a population'. This may seem unambiguous: the population living in an isolated woodland occupies the area of that woodland; the area occupied by a population studied in a sampling grid is the area of that grid; the area occupied by a population of pigs is the area of the pens through which they roam. Implicit in this simple view, however, is the assumption that these populations *fully* occupy the areas concerned. But in practice, many populations are likely to leave areas unoccupied within their boundaries. A doubling of population size within an unchanged boundary may then appear to imply a doubling of density, whereas in reality the density might increase by much less than this or even remain unchanged, because the additional individuals have occupied previously unoccupied parts of the habitat. At its most extreme, this is equivalent to De Jong et al.'s 'characteristic area' for each individual: number and area are directly proportional, density is constant, and transmission is frequency-dependent as a consequence.

THE DIMENSIONS OF β

The various parameters β that we have introduced are different in that they have different dimensions (a point not raised by McCallum et al. [10]). A discussion of these throws further light on their interrelations. The dimensions of the β in equations (2) and (4) (density-dependent transmission) are inherited from those of κ; they are area (per individual) per unit time, reflecting the fact that this latter quantity can be regarded as the effective area over which a susceptible makes contact in unit time. The dimensions of the β' in equations (3) and (5) (frequency-dependent transmission) are inherited from those of η; they are time^{-1}, since this quantity is the rate at which a susceptible makes contact with other hosts. Finally, the β^* in equation (6) (the conventional 'mass action' equation) inherits its dimensions from those of a contact rate per susceptible, which is taken to increase in proportion to the total number of individuals in a population occupying a constant area. The dimensions are thus (per individual) per unit time.

THE BIOLOGY OF TRANSMISSION: 'HOMOGENEOUS AND HETEROGENEOUS MIXING'

What, then, are the biological meanings of density- and frequency-dependent transmission? In the past, answers to this question have often involved implicitly equating density-dependent transmission with 'homogeneous mixing' – random contacts between individuals – and frequency-dependent transmission with some contrasting type of heterogeneity: for example, the selection of partners as an element in the transmission of a sexually transmitted disease [18]. There is, however, no such direct correspondence. If mixing is heterogeneous in the sense that there are distinct classes amongst hosts in terms of their 'contact experience' (for example, young individuals are static, older individuals move freely), then no single transmission term can capture this – separate terms are requited for transmission within and between the different classes.

Homogeneity of contact experience, though, can come about in two ways. First, each individual may have an equal chance of contacting any other individual in the whole population during a given time period, as when individuals move rapidly and at random throughout the population. This is homogeneous mixing in every sense. But, contrary to what is often assumed (e.g. [10]), this is not the only route to homogeneity of contact experience. The contact experience is also homogeneous when all individuals have equivalent contact structures – including rates – (i.e. the system is uniform, at least in some statistical sense). In this case, contact structures can be heterogeneous (e.g. only involving local interactions) without destroying the homogeneity of the contact process. Now, in a uniform system with equivalent contact structures (including rates), we expect on average that the distribution of susceptibles, infecteds, etc. will look the same for each individual – all susceptibles, for example, may experience locally approximately the same (global) prevalence of infection, I/N. This additional feature means that there will be homogeneity of contact experience. (Naturally, local interactions will create local heterogeneities – as when an infectious individual gives rise to a temporary cluster of infected individuals – but this will simply

Table 1. *Illustrative examples of biological scenarios corresponding to four alternative types of transmission*

	Density dependent transmission: density dependent contact rate	Frequency dependent transmission: constant contact rate
Homogeneous contact structure	1. Individuals move throughout the population, fighting others at random. 2. All infectious particles rain down equally (whatever their source) on all susceptibles.	1. Random sexual encounters (once a night but it could be with anyone). 2. Random 'social' behaviour – a fixed frequency of fights, for example, but these could be with anyone.
Heterogeneous contact structure within a homogeneously distributed population	1. Individuals fight their neighbours, the numbers of which are proportional to global density. 2. Infectious particles are dispersed only locally (but the rain of particles is nonetheless homogeneous). 3. Family member contact, where family size is proportional to global density.	1. Non-random sexual encounters (once a night with those closest to hand). 2. Fights with territorial neighbours (where territories change in size with density so that the number of neighbours is always the same). 3. Family member contact, where family size is the same irrespective of density.

create noise around a shared 'expectation' of contact experience.) The important contrast, therefore, is between (i) truly homogeneous contact and (ii) a heterogeneous contact *structure* within a uniform system, nonetheless leading to homogeneous contact *experience*.

This contrast, moreover, is orthogonal to the distinction between density- and frequency-dependent contact rates and transmission. The total per capita contact rate can either scale with density (in the density-dependent case) or remain constant (in the frequency-dependent case) through either a sequence of homogeneous contact structures or a sequence of heterogeneous ones. Illustrative examples (not exhaustive) of the biological scenarios that may lead to the four consequent types of transmission are listed in Table 1. Bearing in mind that these are benchmark transmission modes, to which real examples are unlikely to conform strictly, it is apparent from the table that frequency-dependent transmission is most likely to be associated with a heterogeneous contact structure, as often suggested (the scenarios in the top-right cell of the table are less plausible than those to the bottom-right). Also, density-dependent transmission is more likely to be associated with a homogeneous contact structure than is frequency-dependence (the scenarios in the top-left cell are perhaps more plausible than those to the top-right). But density-dependent transmission will clearly often arise out of a heterogeneous contact structure (lower-left), while frequency-dependent transmission with a homogeneous contact structure is far from unimaginable (top-right).

INCLUDING VARIATIONS IN AREA

The decision to base our account on equations for numbers means that those in densities acquire extra terms in situations where area varies dynamically. Thus equation (2) leads to

$$di/dt = \beta si - (i/A)\,(dA/dt) = i(\beta s - dA/dt \cdot I/A) \quad (7)$$

in place of equation (4). The extra term here may be interpreted biologically as follows. When area is increasing over time, the rate of production of new infecteds is lowered in line with the proportionate increase in area ($dA/dt \cdot I/A$), because the contact rate is proportionately lower.

CONCLUSION

We have concentrated here on reviewing and then clarifying the meaning of two simple transmission terms in common use. As we and others have noted previously, however [2, 14, 19] transmission in practice, even when it can be reduced to a single term, will only rarely conform exactly to either density- or frequency-dependence, instead either lying somewhere between the two, or being better described by some variant more or less closely related to them. McCallum et al. [10] have reviewed the range of alternatives that have been suggested. Nonetheless, for the foreseeable future, the two terms discussed here are likely to remain benchmarks against which actual transmission dynamics are judged, and to remain key elements in most mathematical models of transmission. Uniformity of terminology and unambiguousness of meaning are therefore highly desirable.

ACKNOWLEDGEMENTS

We thank NERC and MAFF for financial support, and Andy Dobson, Rob Knell and Hamish McCallum for their comments on the manuscript.

REFERENCES

1. Grenfell BT, Dobson AP. *Ecology of infectious diseases in natural populations*. Cambridge: Cambridge University Press, 1995.
2. De Jong MCM, Diekmann O, Heesterbeek JAP. *How does transmission of infection depend on population size?* In: *Epidemic models: their structure and relation to data models*. Mollison D, ed. Cambridge: Cambridge University Press, 1995: 84–94.
3. Bouma A, De Jong MCM, Kimman TG. Transmission of pseudorabies virus within pig-populations is independent of the size of the population. Prev Vet Med 1995; **23**: 163–72.
4. Anderson RM, May RM. Population biology of infectious diseases: Part 1. Nature 1979; **280**: 361–7.
5. Smith G, Grenfell BT. Population biology of pseudorabies in swine. Am J Vet Res 1990; **51**: 148–55.
6. Thrall PH, Biere A, Uyenoyama MK. Frequency-dependent disease transmission and the dynamics of the *Silene-Ustilago* host-pathogen system. Am Nat 1995; **145**: 43–62.
7. Fromont E, Pontier D, Langlais M. Dynamics of a feline retrovirus (FeLV) in host populations with variable spatial structure. Proc R Soc Lond Ser B-Biol Sci 1998; **265**: 1097–104.
8. Thrall PH, Antonovics J. Polymorphism in sexual versus non-sexual disease transmission. Proc R Soc Lond Ser B-Biol Sci 1997; **264**: 581–7.
9. De Leo GA, Dobson AP. Allometry and simple epidemic models for microparasites. Nature 1996; **379**: 720–2.
10. McCallum H, Barlow N, Hone J. How should pathogen transmission be modelled? Trends Ecol Evol 2001; **16**: 295–300.
11. Hamer WH. Epidemic disease in England – the evidence of variability and the persistence of type. Lancet 1906; **1**: 733–9.
12. Guldberg CM, Waage P. *Etudes sur les affinities chimiques*. Christiania (Oslo): Brogger and Christie, 1867.
13. Pannetier G, Souchay P. *Chemical kinetics*. New York: Elsevier Publishing Company, 1967.
14. Begon M, Feore SM, Bown K, Chantrey J, Jones T, Bennett M. Population and transmission dynamics of cowpox in bank voles: testing fundamental assumptions. Ecol Lett 1998; **1**: 82–6.
15. Swinton J, Harwood J, Grenfell BT, Gilligan CA. Persistence thresholds for phocine distemper virus infection in harbour seal *Phoca vitulina* metapopulations. J Anim Ecol 1998; **67**: 54–68.
16. Anderson RM. Populations and infectious diseases: ecology or epidemiology? J Anim Ecol 1991; **60**: 1–50.
17. Begon M, Hazel SM, Baxby D, et al. Transmission dynamics of a zoonotic pathogen within and between wildlife host species. Proc R Soc Lond Ser B-Biol Sci 1999; **266**: 1939–45.
18. Lockhart AB, Thrall PH, Antonovics J. Sexually transmitted diseases in animals: ecological and evolutionary implications: Biol Rev Cambridge Philosophic Soc 1996; **71**: 415–71.
19. Antonovics J, Iwasa Y, Hassell MP. A generalised model of parasitoid, venereal and vector-based transmission processes. Am Nat 1995; **145**: 661–75.