Stochastic Processes in the Causation of Rheumatic Disease

PETER J. ROBERTS-THOMSON, MICHAEL E. JONES, JENNIFER G. WALKER, JAMES G. MACFARLANE, MALCOLM D. SMITH, and MICHAEL J. AHERN

ABSTRACT. ***Objective.*** Rheumatic disorders arise in certain individuals depending on the interaction of genetic and environmental factors, the contribution for each varying with the specific rheumatic disorder. However, a third variable, i.e., random or stochastic processes, may be important, but this has been poorly studied. We examined 3 rheumatic disorders to determine whether a simple stochastic process might be consistent with the incidence data.

Methods. A questionnaire and clinical survey of patients with ankylosing spondylitis, rheumatoid arthritis, and systemic sclerosis was performed to determine age at onset of first symptom. Population data were obtained from the Australian Bureau of Statistics. Computer modeling of the equation

$$\frac{dN}{dt} = kP_0 t^{r-1}\exp(-kt^r/r)$$

was performed, where dN/dt is the age-specific incidence rate, P_0 is the proportion of population at risk, t is the age at onset, k is a constant, and r is the number of random events that must occur before the disease manifests.

Results. Analysis of the age-specific incidence for each of these 3 rheumatic disorders was consistent with the stochastic model, where r varied from 4 to 9.

Conclusion. An examination of the age-specific incidence suggests that only a small number of random events need to occur in a predisposed population to allow the emergence of the rheumatic disorder. These random events might be environmental (e.g., infections or exposure to toxins) or due to acquired genetic changes (e.g., somatic mutations involving pivotal immune or growth/repair genes). (J Rheumatol 2002;29:2628–34)

Key Indexing Terms:
AGE-SPECIFIC INCIDENCE SYSTEMIC SCLEROSIS RHEUMATOID ARTHRITIS
ANKYLOSING SPONDYLITIS STOCHASTIC PROCESS

Disease causation is generally ascribed either to genetic factors (e.g., hemochromatosis) or environmental factors (e.g., cigarette smoking, asbestos exposure), or a combination of both (e.g., diabetes) and there is much research in assessing the individual contribution. However, some disorders occur randomly in a population without strong genetic or environmental associations, and one explanation of this random occurrence is to assume that stochastic processes (i.e., random events) have occurred in a predisposed population[1]. Such predisposition, of course, might have a specific genetic basis. These random events could imply exposure to specific infections or toxins, or alternatively might represent random mutations in pivotal somatic genes involved in cellular growth and differentiation, DNA repair, or in immune mechanisms. The essence of the randomness that we envisage here is that genetically similar individuals (even identical twins) might be exposed to what we consider identical environments, and yet the event might occur in one individual and not the other.

We investigated possible random events in disease causation in 3 rheumatic disorders by analyzing the age-specific incidence of these diseases. Analysis of these data revealed that the incidence of these diseases is consistent with a stochastic process. The model itself is not new. Nordling[2], reviewing and extending earlier work, discussed a model of cancer in which a requirement for n independent mutations would give age-specific incidence that increases as the $(n - 1)$ power of age. Armitage and Doll[3] extended the mathematical development of the theory and applied it to several different cancers.

Burch and Rowell[4] applied a similar model to autoimmune disease, and their equation, given below, attributes the

From the Departments of Immunology, Allergy and Arthritis and Anatomy and Histology, Flinders Medical School, Bedford Park, Adelaide, Australia.

P.J. Roberts-Thomson, MD, DPhil (Oxon), Professor and Chairman, Department of Immunology, Allergy and Arthritis; M.J. Jones, MBBS, PhD, Associate Professor, Department of Anatomy and Histology; J.G. Walker, MBBS, Department of Immunology, Allergy and Arthritis; J.G. Macfarlane, Medical Student, Oxford University, Oxford, UK; M.D. Smith, MBBS, PhD, Professor; M.J. Ahern, MD, Associate Professor, Department of Immunology, Allergy and Arthritis.

Address reprint requests to Prof. P.J. Roberts-Thomson, Department of Immunology, Allergy and Arthritis, Flinders Medical Centre, Bedford Park SA 5042, Australia.

Submitted November 20, 2001; revision accepted June 13, 2002.

time of onset of the disease to the time at which the last of several mutations takes place.

$$\frac{dN}{dt} = kP_0 t^{r-1}\exp(-kt^r/r) \quad (1)$$

where dN/dt is the age-specific incidence rate, P_0 is the proportion of population at risk, t is the age at onset, k is a constant, and r is the number of random events that must occur before the disease manifests.

This age-specific incidence rises to a peak and then falls, whereas the earlier models relating to cancer increase with time. Mathematical details of the equations and their derivation are given in the Appendix.

MATERIALS AND METHODS

Patients. Patients with ankylosing spondylitis (AS) and rheumatoid arthritis (RA) were identified from the disease indexes of the Rheumatology Units at Flinders Medical Centre, Repatriation General Hospital, and Queen Elizabeth Hospital, Adelaide, Australia. Patients with progressive systemic sclerosis (SS) we identified from 2 sources: (1) the South Australian Scleroderma Registry[5], a register of all known scleroderma patients resident in South Australia, and (2) the Sydney Scleroderma Epidemiological Survey coordinated by Dr H. Englert[6,7]. Diagnosis in each patient was by the attending rheumatologist based on clinical, laboratory, and radiological features and according to standard American Rheumatology Association criteria[8] in the cases of RA and SS. In addition, for scleroderma, patients were entered if they had sclerodactyly plus 2 or more of the features of Raynaud's phenomenon, esophageal dysfunction, calcinosis, telangiectasia, nailfold capillary abnormalities, or antinuclear antibodies in order to include all subsets, as reported[5-7]. Diagnosis of AS was dependent upon the presence of appropriate clinical features together with radiological evidence of sacroiliitis and spondylitis. A summary of demographic, clinical, and serological features of the 3 disease groups is given in Table 1.

Age at disease onset. Age at disease onset was defined as the age at initial symptom of the rheumatic disease. This was obtained from (1) circulation of a mailed questionnaire to each patient asking them to date their age at first symptom, or (2) in the patients with scleroderma from the date on file in the disease indexes (originally obtained from a previous questionnaire survey[5-7]).

To confirm validity of the age of disease onset obtained from the questionnaire with disease onset as determined from clinical or patient interviews, a subgroup of 100 patients with scleroderma was compared for whom this information was available from both sources (i.e., the questionnaire or clinical notes).

Computing. Mathematical details appear in the Appendix. Briefly, a suite of programs were developed based on the stochastic model underlying equation (1) above. A Monte Carlo simulation was constructed to generate simulated data in which population size, susceptibility, number of random events, and the probability per unit time of an event occurring could be varied. A maximum likelihood method was used in an analysis program to determine the parameters of best fit from a data set. The analysis program was validated against a variety of simulated data sets before applying it to the real clinical data. Population data for South Australia were obtained from the Australian Bureau of Statistics. The point prevalence for pSS in South Australia is 0.023%[5], while a point prevalence of 1% was assumed for RA and 0.1% for AS[9]. The incidence of pSS is 1/15 of the prevalence[5], while an incidence of 1/10 the prevalence was assumed for RA, and an incidence of 1/30 of the prevalence was assumed for AS[9].

Table 1. Demographic, clinical, and serological features of rheumatic disease groups.

Ankylosing Spondylitis	
M:F	78:124
Mean disease duration, yrs (range)	25.2 (2–76)
% Uveitis	49
% Colitis	16
% Peripheral arthritis	62
Rheumatoid Arthritis	
M:F	91:149
Mean disease duration, yrs (range)	16.8 (1–60)
% Seropositive	73
% Positive for shared rheumatoid epitope	75
Health Assessment Questionnaire score (± SD)	1.01 ± 0.84
Systemic Sclerosis	
M:F	222:705
Mean disease duration*, yrs	16.4 (1–69)
Limited: diffuse	718: 209
% ANA positive*	94
% Anticentromere positive*	51
% Scl–70 positive*	11

* South Australian cohort data only.

RESULTS

Figure 1 shows the theoretical curves obtained for age-specific incidence plotted against age at onset for a disease due to stochastic processes. The outcome of a Monte Carlo simulation is shown together with the expected age-specific incidence for the same parameters as predicted by equation (1). It was observed that the shape of the curve was dependent on r, the number of random events and μ, the proba-

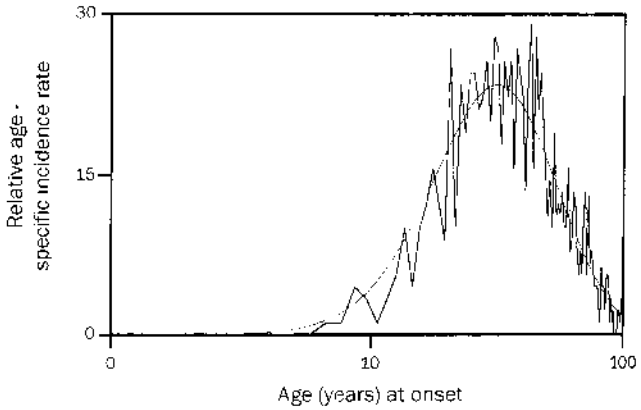

Figure 1. Theoretical curves obtained for age-specific incidence/age at onset for a disease due to stochastic process, where r = number of random processes = 5 and μ = random event rate/year of 0.05 have been assumed. Continuous line: theoretical curve obtained from deterministic equation; jagged line: theoretical curve obtained from Monte Carlo trial.

bility (per unit time) of an event occurring; increasing r gives a sharper peak, while the peak can be delayed by increasing r or decreasing μ. Obviously the height of the curve depends on P_0, the proportion of the population at risk.

Clinical validation that the questionnaire analysis of age at disease onset gave results consistent with the clinical data in 100 patients with scleroderma is illustrated in Table 2.

The best-fit curves for age-specific incidence/age at onset derived for AS, RA, and pSS (subdivided into the specific curves for sex and for limited and diffuse scleroderma) are shown in Figures 2 and 3. In addition these figures indicate the original data as a series of points about this curve, together with the 95% confidence limits within which one would expect the data point to lie. The parameters of best fit are given in Table 3.

The sensitivity analysis examines the confidence regions for the model parameters. These vary from relatively wide, in the case of AS in females, for which there are few cases on which to base the parameter estimates (Figure 4), to narrow for scleroderma, for which there are far more data. The 50% confidence regions for the 8 subgroups are shown in Figure 5. From Figure 4, it can be seen that the 95% CI contour does not overlap zero, and further, in each of the other subgroups in no instance does this 95% CI contour cross zero. This observation therefore enables the conclusion to be made that the null hypothesis for zero stochastic events for our model can be refuted. Details of the calculations appear in the Appendix.

DISCUSSION

From data given in Table 3, it can be concluded that only a small number of random events need to occur in a predisposed population to allow the emergence of the clinical disease (pSS, RA, or AS). Our results do not allow us to determine the nature of these random events, but they could be either environmental exposures to possible infections or toxins, or due to mutations in critically important genes in cellular growth and differentiation, DNA repair, or immune responses to newly mutated autoantigens[10], or alternatively, non-germline genetic variations such as T cell receptor and immunoglobulin gene rearrangement and methylation patterns[1].

It is important to note that a number of assumptions have been made in determining the frequency and number of the random events. These include the following: (1) There is a constant rate of stationary random events (i.e., there is equal probability as a function of time). In reality, it is more likely that there is a changing probability with time. For instance, flying at high altitude increases the risk of background irradiation, and a changing probability could account for a change in the shape of the stochastic curve. (2) The order in which the random events occurs is not relevant. Again, in reality, the order may be very relevant. However, this would not necessarily lead to any change in the shape of the curve. (3) All events are equally probable.

The data are close to, but deviate from, the model. The model is, of course, unreasonably constrained in that all events have equal probability of occurring, and that probability is constant throughout time. Both of these assumptions are unreasonable; even somatic mutations have different probabilities depending on where they are in the genome, and mutation rates may vary with time. Certainly the deviations of the data from the model go beyond those that can be attributed to chance; the model is too simple to reflect the finer points of the data. There is little to be gained, however, by an "ad hoc" fitting of events with differing probabilities or by letting the probabilities vary with time. Indeed, a single random event, the probability of which varies appropriately with time, can model the age-specific incidence of any disease. Such a model is no more an "explanation" of the disease than is an appeal to bad karma or an evil spirit in the sky. But it is just as unscientific to deny the relevance of random events as it is to attribute everything to "fate." Unravelling the role of stochastic events in disease etiology will involve treading a very fine line.

There are 2 plausible sources of "delay," neither of which has been added to the parameters of the model. One is diagnostic delay; there must be a time between occurrence of the last "event" and expression of the disease, and also between that and formal diagnosis. It is likely that the diagnostic delay would be age-dependent. The second delay relates to

Table 2. Comparison of age at onset obtained from questionnaire and clinical notes. Data are mean age at onset [years (± SE)].

	Clinical Notes	Questionnaire
Diffuse, n = 28	47.7 (2.42)	48.1 (2.46)
Limited, n = 61	44.1 (1.85)	43.2 (2.06)
Overlap, n = 11	33.5 (4.94)	32.5 (4.95)

Table 3. Calculated stochastic variables obtained for patients with scleroderma, RA, and AS.

	Women				Men			
	Patients (n)	μ	r	Po	Patients (n)	μ	r	Po
Limited scleroderma	567	0.03	5	0.003	151	0.02	6	0.001
Diffuse scleroderma	138	0.04	8	0.0005	71	0.03	8	0.0004
RA	149	0.02	4	0.33	91	0.02	6	0.17
AS	24	0.08	7	0.001	78	0.09	9	0.002

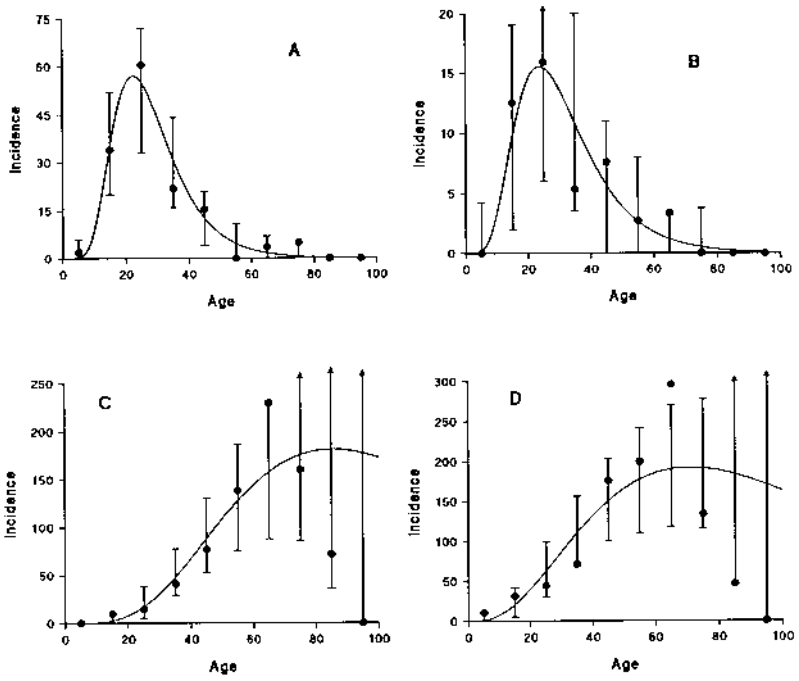

Figure 2. Age-specific incidence rate plotted against age at onset from first symptoms for AS and RA. Bars represent 95% confidence intervals for the best-fit curve. A: males with AS; B: females with AS; C: males with RA; D: females with RA. The y ordinate = incidence = number of new cases/year per 1×10^4 population.

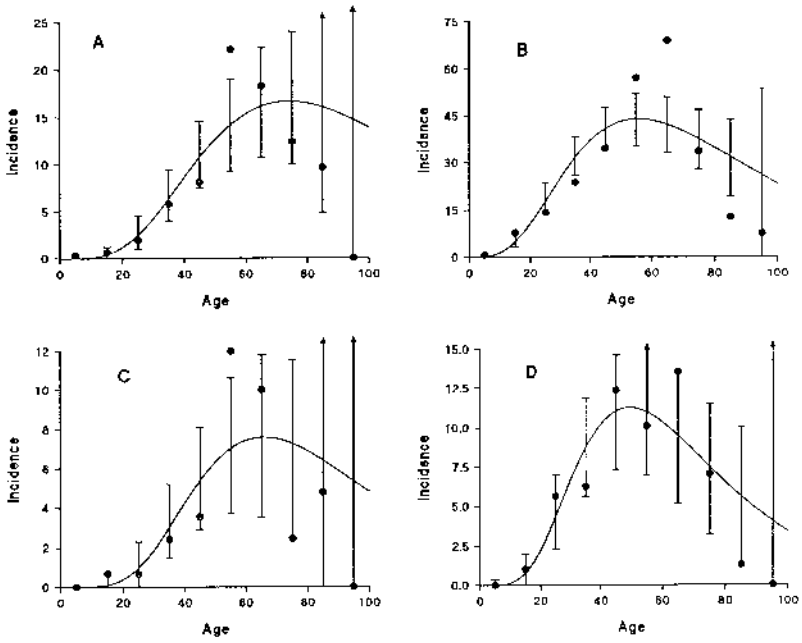

Figure 3. Age-specific incidence rate plotted against age of onset from first symptom for scleroderma. Bars represent 95% confidence intervals for the best-fit curve. A: males with limited scleroderma; B: females with limited scleroderma; C: males with diffuse scleroderma; D: females with diffuse scleroderma. The y ordinate = incidence = number of new cases/year per 1.5×10^5 population.

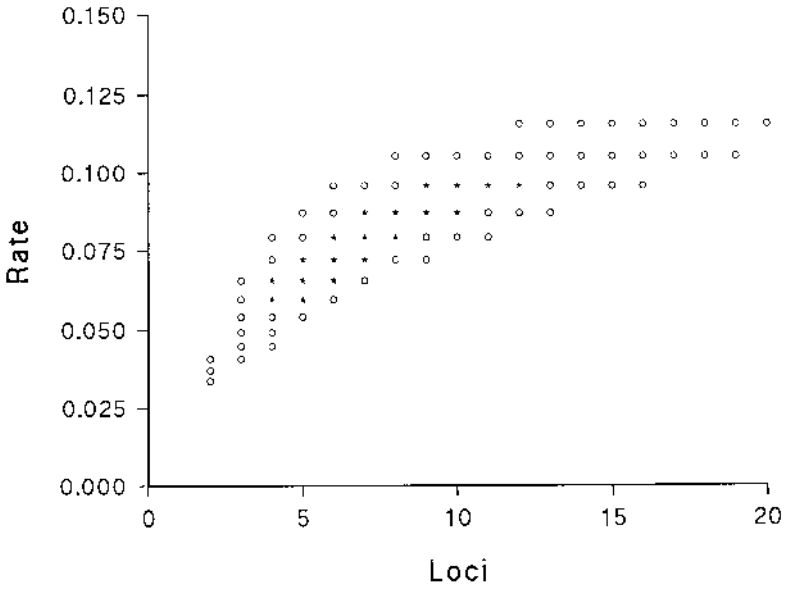

Figure 4. Confidence regions for parameter estimates are shown for AS in females. The centrally placed 50% confidence interval is illustrated as ★, about which the 95% confidence interval is shown as ○.

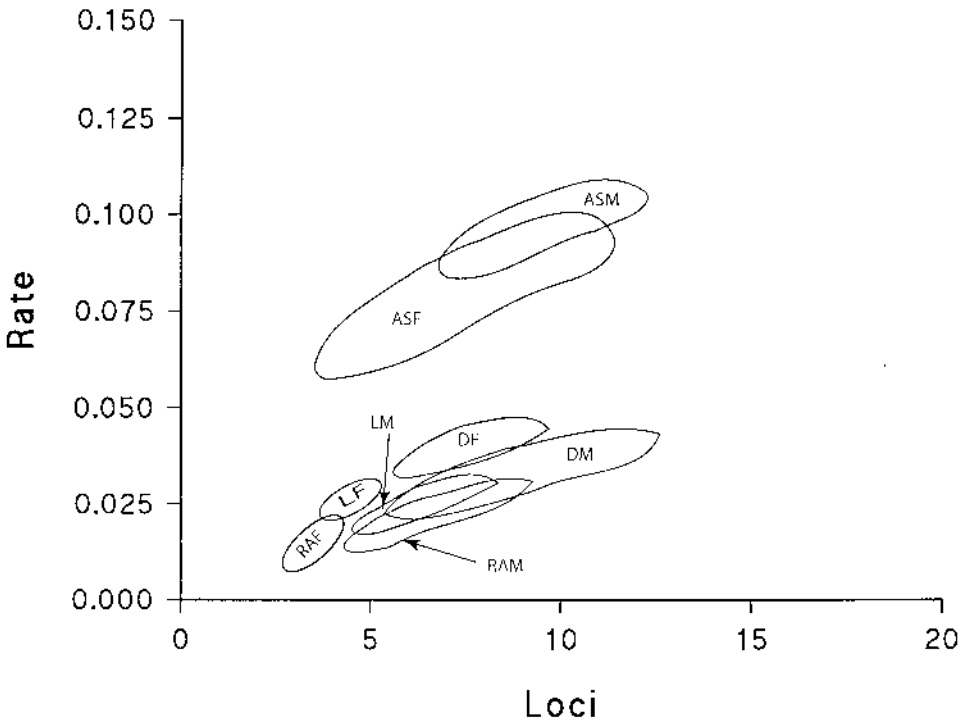

Figure 5. Fifty percent confidence interval contours for the 3 rheumatic diseases; RAM: rheumatoid arthritis in males, RAF: RA in females, ASM: ankylosing spondylitis in males, ASF: AS in females, LM: limited scleroderma in males, LF: limited scleroderma in females, DM: diffuse scleroderma in males, DF: diffuse scleroderma in females.

the difference between incidence and prevalence. In a disease such as RA, where the incidence is a tenth of the prevalence, average survival after diagnosis is 10 years, and there is a plausible argument that prevalence most closely reflects the incidence of about 5 years previously. But survival, too, is likely to be age-dependent. We have decided against allowing for these factors by adding further parameters to the model. We have retained only the 3 parameters proposed by Burch and Rowell[4].

From a consideration of the above, it is clear that one

may evoke these caveats in accounting for the actual curve deviating from the theoretical curve as shown in Figure 1. The surprising thing is the extent to which this simplistic model of a small number of loci or events with equal and constant "hazard" models the observed data in these 3 rheumatic disorders and may well be applicable to other human disorders as well.

What is the possible relevance of our stochastic modeling in explaining the incidence of these rheumatic diseases? Our models support the possibility that a small number of stochastic events is consistent with and supportive of the observed incidence rates. Such a conclusion is important because, as described by Gregersen[1], it may explain the relatively low monozygotic twin concordance rate in diseases with high heritability (e.g., RA, where a recent study reports a concordance rate of 0 in monozygotic twins and 8.8 in dizygotic twins[11]). Further, it may also explain the incidence of diseases such as scleroderma, where detailed studies have revealed no strong genetic or environmental linkage[5].

A recent viewpoint article in *Lancet* has discussed nature, nurture, and stochastic processes in the causation of complex human disorders such as RA[12]. Stochastic variables were defined as environmental, genetic, or non-attributable variables and were felt to be important in the etiology of these complex disorders. We believe our modeling is consistent with the likelihood of stochastic factors being operative in such disorders, and our findings give some definitive data as to their number and frequency.

An examination of the age-specific incidence/age of onset relationship in 3 rheumatic diseases suggests that only a small number of random events (numbering from 4 to 9) need to occur in a predisposed population to account for the emergence of the disorder. The identity of these random events is unknown, but might include mutations in pivotal genes involved in cellular growth or regulation, DNA repair, or immune responses.

ACKNOWLEDGMENT

We acknowledge Dr. K. Pile, Queen Elizabeth Hospital, in recruitment of patients with ankylosing spondylitis and M. Emin in secretarial assistance.

APPENDIX

Historically, the literature relating to the accumulation of random events in the etiology of disease has focused on 2 models that are initially difficult to reconcile. Nordling[2] argued that the incidence of many cancers increases as the 6th power of age, and that this is consistent with there being 7 successive mutations. In his model the incidence increases without limit, and he emphasizes that, should incidence ultimately decline, then the hypothesis of cumulative mutations must be rejected.

In apparent contrast, the equation that Burch and Rowell[4] use to model the incidence rises to a maximum and then declines, an observation that Nordling suggests should lead us to reject the hypothesis of cumulative mutation.

We derive, and reconcile, the equations below.

If μ_i is the probability per unit time of the mutation i occurring, then the probability that it has not occurred at time t is $e^{-\mu_i t}$. Then $P_i(t)$, the probability that it has occurred, is given by

$$P_i(t) = 1 - e^{-\mu_i t}$$

Providing $\mu_i t$ is small, this can be very closely approximated by

$$P_i(t) \approx \mu_i t$$

Using that approximation, consider a disease that requires r mutations. At a given time, the proportion of the population "on the verge" of developing the disease are those who have already undergone $(r-1)$ mutations and are awaiting the occurrence of the r^{th} mutation. The probability that any one individual is in that state at time t is

$$P_1(t) \times P_2(t) \times \cdots \times P_{r-1}(t) = \mu_1 \times \mu_2 \times \cdots \times \mu_{r-1} t^{r-1} = kt^{r-1}$$

for some constant $k = \mu_1 \mu_2 \cdots \mu_{r-1}$. Depending on the presence or absence of constraints on the order in which particular mutations must occur, k can be larger or smaller, but time still enters the equation raised to the $(r-1)^{th}$ power.

If we now multiply the proportion who have undergone the $(r-1)$ mutations by the rate of undergoing the final, r^{th} mutation, we have the incidence

$$I(t) = kt^{r-1}\mu_r$$

This, the age-specific incidence, rises as the $(r-1)^{th}$ power of age, if r mutations are involved. The approximations involved in this derivation are that the individual probabilities, $\mu_i t$, are small, so that $e^{-\mu_i t} \approx 1$ and $1 - e^{-\mu_i t} \approx \mu_i t$. As t increases, the approximation becomes less acceptable; not only does the event become certain, but the model assumes that the disease may occur on multiple occasions. One route from Nordling's model in which the age-specific incidence increases without limit, to Burch's model in which there is an observed peak, is to consider only the first incidence, in any one individual, of the disease in question.

If the age-specific incidence of a disease that can occur on multiple occasions is given by

$$I(t) = kt^{r-1}$$

then the expected number of incidences, $M(t)$, in any one individual, up until time t, is given by

$$M(t) = \int_0^t I(t)\, dt$$

$$= kt^r/r$$

Following a standard result on Poisson processes, the probability, $P_0(t)$, that the disease has not developed, *(i.e.,* has developed zero times), at time t is

$$P_0(t) = e^{-M(t)}$$

and the probability, $I_1(t)$, that it has its first occurrence at time t is then

$$I_1(t) = -\frac{dP_0(t)}{dt}$$

$$= e^{-M(t)} \frac{dM(t)}{dt}$$

$$= kt^{r-1} e^{-kt^r/r}$$

which is the equation used by Burch.

$I_1(t)$, as expressed here, is the probability that a single susceptible individual will develop the disease at time t. In a population of size N, of which a proportion s are susceptible, the age-specific incidence will be $NsI_1(t)$.

Implementation on a computer

Deterministic model. The above uses the classical approach using calculus and the theory of Poisson processes to reconcile the equations of Nordling and of Burch. In implementing the model on a computer we have taken the liberty of going directly to a discrete approach, thus largely bypassing the need for approximations, calculus, and Poisson processes. The underlying model is unchanged; there are r events that must occur, in any order, and the disease develops in the year in which the last of these takes place. The algebra is straightforward, and is perhaps more accessible than the mathematics above.

Let μ be the probability that a given random event occurs in a year. Then $(1 - \mu)$ is the probability that it does not occur in a year, and $(1 - \mu)^n$ is the probability that it has not occurred by the end of the n^{th} year. The probability that it has occurred by the end of the n^{th} year is $[1 - (1-\mu)^n]$. If there are r such events, then the probability that all of them have occurred at the end of the n^{th} year is $[1 - (1 - \mu)^n]^r$. Finally, the probability that this state develops in the course of the n^{th} year is the probability that it is so at the end of the n^{th} year, minus the probability that it was so at the end of the previous year.

Accordingly, $I(n)$, the probability of the disease arising in a given individual in the n^{th} year is given by

$$I(n) = [1 - (1 - \mu^n]^r - [1 - (1 - \mu)^{n-1}]^r$$

Stochastic model. All the preceding relates to a stochastic model, for which we have written essentially deterministic equations. To model the randomness we have programmed a Monte Carlo simulation. Each person in the population is simulated as an array of r loci, none of which are mutated at time zero (birth of the individual). Using a random number generator, each unmutated locus has probability μ of mutating each year. If, at the end of 100 years there remain unmutated loci, then the simulated person is assumed to have lived without developing the disease. If, however, all r loci mutate, then the year of the last mutation is the year in which the person develops the disease.

Parameter estimation. The models above address the "forward" problem; with a given number of loci and a given mutation probability, we calculate a predicted outcome. The "inverse" problem of estimating numbers of loci and mutation probabilities from available clinical data uses the classical techniques of maximum likelihood estimation. Standard mathematical subroutines published in connection with Press, *et al*[13] were used for these calculations. For the age-specific incidence data of Figures 2 and 3 the confidence limits shown reflect both the populations size and the predicted incidence. Thus, for instance, the model predicts an age-specific incidence for RA in males between the ages of 50 to 59 years of 129 per 100,000. This corresponds, in our observed population, to an expectation of 18.7 occurrences. A Poisson variable with this parameter will take the integral values of 11 to 27 inclusive 95% of the time, corresponding to an observed age-specific incidence per 100,000 of 76 to 187, as indicated by the error bars. Where the observed population is small, even one occurrence, although lying within the 95% bounds, corresponds to an "off scale" incidence and these occurrences are signified by upward arrows replacing the usual error bars.

Confidence limits for parameter estimates in Figures 4 and 5 are derived using a Bayesian approach assuming equal priors. Clearly, the size of the confidence area depends on the size of the data set. Thus, bounds for RA in females at one extreme are much tighter than for ankylosing spondylitis in females at the other.

REFERENCES

1. Gregersen PK. Discordance for autoimmunity in monozygotic twins. Arthritis Rheum 1993;36:1185-92.
2. Nordling CO. A new theory on the cancer-producing mechanism. Br J Cancer 1953;7:68-72.
3. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954;8:1-12.
4. Burch PRJ, Rowell NR. Autoimmunity: aetiological aspects of chronic discoid and systemic lupus erythematosus, systemic sclerosis and Hashimoto's thyroiditis. Lancet 1963;2:507-13.
5. Roberts-Thomson PJ, Jones M, Hakendorf P, et al. Scleroderma in South Australia: epidemiological observations of possible pathogenic significance. Intern Med J 2001;31:220-9.
6. Englert H, Small-McMahon J, Chambers P, et al. Familial risk estimation in systemic sclerosis. Aust NZ J Med 1999;29:36-41.
7. Englert HJ. The epidemiology of scleroderma in Sydney, 1974-1988: A comparative study [thesis]. Sydney: University of Sydney; 1993, 351.
8. Klippel JH, editor. Primer on the rheumatic diseases. Appendix 1. Atlanta: Arthritis Foundation; 1997:453-64.
9. Masi AT, Medsger TA. Epidemiology of the rheumatic diseases. In: McCarty DJ, editor. Arthritis and allied conditions. A textbook of rheumatology. 2nd ed. Philadelphia: Lea and Febiger; 1989:16-54.
10. Bachmann P, Semsei I, Gross JD, Gross TF, James JA, Harley JB. Somatic mutation: a novel mechanism for initiation of autoantibodies [abstract]. Arthritis Rheum 2001;44 Suppl:S273.
11. Svendsen AJ, Holm NV, Kyvik K, Petersen PH, Junker P. Relative importance of genetic effects in rheumatoid arthritis: historical cohort study of Danish nationwide twin population. BMJ 2002;324:264-7.
12. deVries N, van Riel PLC, van de Putte LBA. Research in complex disease. Lancet 2002;359:1243-45.
13. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes; the art of scientific computing. Cambridge: Cambridge University Press; 1988.