SOME EMPIRICAL STUDIES IN EPIDEMIOLOGY

C. C. SPICER
GENERAL REGISTER OFFICE, LONDON

1. Mathematical epidemiology

1.1. *Introduction.* The mathematical theory of epidemics has succeeded in producing a rather satisfactory account of the behavior, in large groups, of certain highly infectious diseases which confer a lasting immunity and have a low proportion of subclinical cases. Measles is the classical example but Professor Bartlett has argued convincingly for the same mechanism in some other acute fevers. These diseases are also characterized by an apparently high stability in the immunological structure of the causative organism.

At the other end of the scale, in small groups, it has been possible to give good accounts of the distribution of epidemic size in households or small communities. In particular, the mathematical concept of a threshold size above which the disease tends to affect the whole group and below which the epidemic peters out without affecting more than a fraction, fits in well with practical experience.

The mathematical theory of geographical spread of epidemics seems to me less well established, because of the lack of good observational data and the complexity of structure of even the simplest types of human community. In a subsequent paper Dr. Bailey [1] discusses the possibility of extending our knowledge in this direction.

1.2. *Problems which might be attacked by mathematical epidemiology.* It seems to me that it is worthwhile considering briefly some problems confronting the public health worker at the moment to whose solution mathematical epidemiology might make a valuable contribution. As I see it, the most important of these in highly developed countries with a temperate climate is presented by the acute viral infections of which influenza is the most important example. In the underdeveloped tropical countries a variety of viral bacterial and protozoal infections are of greater importance but good statistical data on them are very scarce. It is, though, worth noting that mathematical analysis has been found of value in the planning and assessment of malaria eradication campaigns when combined with adequate field work for the collection of basic data. The outstanding characteristics of the acute viral infections which any theory of their behavior must take into account are:

(a) the great antigenic lability and diversity of the viruses;

(b) frequently, the rather short period of immunity that follows an attack;

(c) a very strongly marked seasonal distribution whose dependence on meteorological factors is undoubted but very poorly understood.

The success of the stochastic differential equations in accounting for the behavior of measles suggests that a modified form, taking into account (b) and (c), would be worth considering and easy to set up. One hopes that their analysis, by a happy blend of empiricism and mathematics on the lines of Professor Bartlett's attack on measles, would yield equally interesting results. However, the existence of antigenic lability introduces a complication which has, so far as I know, not been considered in detail and might wreck the chances of practical application of mathematical theory.

Influenza provides a well understood example here of a situation which probably exists for many other types of virus. The sudden appearance of antigenic variants like the Asian type in 1957 seems to be of an essentially unpredictable nature. If the example of Asian flu is typical, the new variant more or less displaces the previous epidemic types and one imagines that it will continue to reign until it is itself displaced by another variant. In the case of the common cold the variants probably follow one another much more quickly and confusingly.

A great variety of agents causing minor illness of influenzal type has now been isolated though the volume of ill health that they cause is not very accurately known. A study of the death rates, notifications of influenza, and claims for sickness benefit under social insurance in England and Wales suggests fairly strongly that the main causes of winter epidemics, which are the economically important sources of morbidity, are the influenza viruses. Study of the records of the past fifteen years gives evidence of the existence of other epidemics of lesser importance, and of quite unidentified cause. It also appears that there is some regular periodicity in the rise and fall of death and sickness rates from these causes and I have made an attempt to extract some evidence from the records of such periodicities. The justification for such a search is that most epidemiologists feel that there is a roughly two year oscillation in the incidence of influenza. It is reasonable to expect that the characteristics of the virus which determine its transmissibility and infectiousness would be more or less independent of the variations in its antigenic type, and consequently a change of type would not alter the basic periodicity of about two years. There might well of course be a change of phase since new types occur at unexpected times of the year, such as Asian influenza in the early autumn and 1918 influenza (if it was influenza) in the summer.

The possible ways of looking for periodicity are rather numerous since different age groups are differently affected and even in a small country like England there is a good deal of geographical heterogeneity. It is, furthermore, quite clear that over the 120 or so years for which records exist in England the determinants of the death rate have varied greatly. During the 19th century the respiratory disease death rate was dominated by tuberculosis, and at the same time there

were other rampant infectious diseases with well marked periodicity such as diphtheria and scarlet fever. Over the whole period there has been a well marked trend in the death rate with a gentle rise during the 19th century followed by a fall which has continued in a roughly exponential fashion into the 20th century and has been maintained to the present day.

1.3. *Spectral analysis approach and its difficulties.* My own approach to these problems has been the rather simple (possibly simple minded) one of removing the trend by fitting polynomials of the second degree separately to the 19th and 20th centuries and then performing spectral analyses on the deviations from the polynomials in the two periods 1838–1890, 1890–1955. The degrees of freedom are not numerous and the width of window available rather restricted so the fine detail of the spectrum is not well brought out. In order to avoid the complications introduced by other infectious diseases, I have confined the analysis to the age groups 45 to 64, and 65 to 74, since these diseases fall mainly in childhood and early life. At the older ages the general death rate and the respiratory disease death rate mirror one another very closely, especially in recent years. The details of the spectral analysis used are not, I think, of major importance since the basic findings are not really affected by using the simple differencing technique given by Tukey and Blackman [5].

The results for the 19th and 20th century are, not unexpectedly, somewhat different. In the 19th century also there are differences in the spectra for males and females which undermine one's confidence in the analysis. However, both spectra and also the separate spectra for the two age groups all agree in showing a maximum of power at a period of two to three years, which falls irregularly as the period increases. There is a secondary peak at around four years in the 20th century and five years in the 19th. In the 20th century there is a small peak at about eight to ten years which is quite absent in the 19th century.

These spectra are in considerable contrast to some others that I have studied in this field. Measles, for example, which has an obvious periodicity of about two years gives, effectively, a line spectrum with a bright line at this period. Leukemia deaths (over a much shorter time) and the birth rate show little evidence of oscillations longer than one year though both show very clear short period fluctuations which for the birth rate take the form of a well marked line spectrum of a fundamental with harmonics.

It is exceedingly tempting to interpret the spectrum as consisting of periods for the two main types of influenza, A and B, since it looks as though influenza A has a period of about two years and influenza B one of four years. Superimposed on this is perhaps a general autoregressive process and a mass of periodicities of smaller amplitude arising from the known multiplicity of acute virus infections. Unfortunately, it is exceedingly difficult to obtain direct evidence for this rather attractive hypothesis.

The main difficulty is that the spectrum gives no information on phase so that it is impossible to identify from it observed fluctuations with known out-

breaks of infection. It is also unfortunate that the main periodicity is so close to the Nyquist frequency and is undoubtedly affected by the occurrence of early and late epidemics around the turn of the year.

1.4. *Use of filters.* In an attempt to take the analysis a little further I have resorted to the use of filters, in the form of simple symmetrical moving averages. The ones I have used were put forward by Craddock [3] some years ago and used by him on meteorological data. When applied to the time series of deviations from trend, these should show the actual positions of the peaks and troughs of the corresponding periodicities. These filters are of course not perfect but are band pass filters which is an advantage in some ways as the exact periods are not known.

For the 20th century the longer eight to ten years filters show quite a well marked but rather erratic wave form which corresponds to no identifiable epidemic disease so far as I can make out from existing records. The four to five year filters show a reasonably well defined standing wave whose later peaks, when records were available, do correspond roughly to known outbreaks of influenza.

The two year filters give results which looked interesting to me, but which expert opinion is inclined to think are either fallacious or at least, misleading. The appearance of the filtered series is that of what are known among electroencephalographers as "spindles," which resemble oscillations in an amplitude modulated carrier wave. It is probable that since one is operating so close to the folding frequency these are complete artifacts but it is possible that they may represent the response of the filter to an impulse or a narrow "box car" function in the basic data such as would be provided by an exceptionally large epidemic. It certainly seems to me that the problem is worth pursuing further by more accurate and sophisticated methods. It would be most interesting to show a continuity of influenza infection over the past sixty years and it would be particularly interesting to see whether the 1918 pandemic could be identified with a known type of virus. In my analysis it coincides neither with flu A or B peak.

A more valuable byproduct of a successful analysis of the periodicity of influenza would be the ability to make more rationally based predictions of the probability of an epidemic. This would allow immunizing measures to be taken in good time though, I suppose, if these were sufficiently effective and energetic they would, in the end, destroy the basis of the prediction.

2. Meteorological factors

2.1. *Influence of meteorological factors.* I would like to turn now to another aspect of epidemiology which is of particular interest in the respiratory diseases, but also in a number of other acute virus infections. This is the influence of meteorological factors on incidence. It has always seemed to me that virtually the whole of the published work on this subject that I have seen is undermined

by two major statistical errors. One is the ignoring of the inevitable correlations between quantities oscillating regularly with about the same period. The other is the ignoring of the very high correlations (by biological standards) between meteorological variates. As regards the first of these, it is only necessary to search among the observed meteorological elements to find one which corresponds, approximately, in phase with the disease in question. Many workers have then assumed on more or less plausible grounds that there is a causal relation between this element and the disease. This is of course a very old form of statistical fallacy and is even further compounded in meteorology because of the high intercorrelations between the weather variables. For example, in any area with a fairly well watered surface a rise in temperature is automatically accompanied by a rise in the absolute humidity and, in fact, both have been suggested as responsible for the summer incidence of poliomyelitis. Again, any fairly long standing temperature inversion in London is accompanied by a rise in atmospheric pollution, and disentangling the effects of these two is very difficult.

Another complication that arises in work on infectious disease is that the strongest influence determining the incidence in any one period is the incidence in the immediate and intermediate past, and it is important to allow for this also. Furthermore, the effects of weather may themselves only show up after a considerable time lag.

I have made several attempts to analyze this situation and avoid some of its pitfalls. The method I have used is naive and certainly oversimplified, but I should like to take this opportunity of putting forward the general problem as worth serious consideration by those with more mathematical expertise.

What I have done in effect is a multiple analysis of covariance on two way tables in which one margin represents years and the other shorter subdivisions of the year, weeks or months. One removes, by standard methods, the main effects of years and, for example, months and then studies the correlations, both contemporaneous and lagged, between the residuals. This has two main advantages: it reduces the often extreme nonorthogonality of the weather variables; and, dealing only with small short term variations, it makes a linear additive model much more plausible and effective. Its greatest disadvantage is that it throws away all the extensive and potentially valuable information provided by the main seasonal swing.

My original analysis was based on a simple approach of this kind and assumed a parallelism of the regression lines of polio on meteorological factors in all months of the year. This was a considerable oversimplification and led me to underestimate the general effect of weather variables, particularly relative humidity, but the effect of temperature was very clearly demonstrated. This work has been criticized (rightly) by Wise [6] who has now made a much more thorough analysis which is to be published shortly. I think it is probably worth giving this in some detail as an illustration of the sort of difficulties that are encountered in this field. I need hardly add that the volume of calculation

necessary would have been quite impossible in the precomputer age. It now seems to me well worth undertaking on a large scale for any disease in which there seems to be a seasonal influence.

The data used both by Wise and myself for the analysis of weather and poliomyelitis were the monthly figures for England and Wales for the years 1947–56. It is more or less essential to work with the logarithm of the poliomyelitis incidence owing to the wide variation in number of cases through the year, the effect of the transformation being to stabilize the relative variability for each month. It was quite clear from the preliminary analysis that lagged regressions on the meteorological variables of up to two months were necessary. It also appeared that the effects of absolute vapor pressure were negligible after allowing for variation in temperature and relative humidity and the analysis was confined to these two latter variables. It is a characteristic of this kind of work that even these two elements may in fact only be standing in for others that were the effective agents.

The method then used by Wise was to fit, for the ten years available, separately for each month, regressions of the form

$$Z = M + \alpha Z_{-1} + GT + cH, \tag{2.1}$$

where Z is the current polio incidence (logarithm), Z_{-1} is the previous month's polio incidence, α, M, b, c, are constants, and T and H are weighted means of the current and previous temperature and relative humidity of the form

$$\begin{aligned} T &= T_{-1} + \lambda_0 T_0 + \lambda_2 T_{-2}, \\ H &= H_{-1} + \lambda_0 H_0 + \lambda_2 H_{-2}; \end{aligned} \tag{2.2}$$

values of λ_0 and λ_2 were determined for the whole year but the regression coefficients b and c were estimated separately for each month. The λ were taken as the same for both of the two meteorological variables. This model is still obviously oversimplified though considerably less so than the one I used. However, it demonstrates quite clearly a very strong association between poliomyelitis, temperature, and relative humidity, especially in the months of July, August, September, and October where they account for between 57 and 90 per cent of the variation of incidence, the effect of temperature being stronger than that of relative humidity.

It would be desirable to complete the analysis by seeing whether the effects at different times of the year were all part of a single, probably nonlinear, relationship between polio incidence and weather. The simple linear relationship which I obtained is, in fact, sufficient to explain the total amplitude of the poliomyelitis oscillations but differs slightly in phase. Wise has suggested that a simple power law would be adequate. It seems to me that the next essential step is to incorporate this into the model, and this would have the additional advantage of recovering some of the information thrown away along with the seasonal constants.

The constant α in (2.1) is of some interest as it represents the "transmissibility"

of the disease. Wise found that a single value for every month of the year was adequate. In other words, the transmission of the disease occurs with about the same readiness at all times of the year and all stages of the epidemic, after meteorological variations have been allowed for. This is an interesting and not by any means obvious finding.

In collaboration with Dr. Walter Holland, I applied my simpler and less accurate analysis to respiratory disease. We have analyzed the emergency admissions to hospital in London for respiratory disease in relation to five meteorological variables and atmospheric pollution. There is a significant relation at ages over 15 years with both atmospheric pollution and temperature (inversely). A later analysis with more data confirmed this and also showed that in infants aged 0 to 4 years there was a relation with temperature but not pollution. I think it is likely that these positive findings will stand, but the negative findings, especially with regard to relative humidity, are less certain and a more thorough analysis will be required.

The most urgent necessity for respiratory disease now would be to obtain estimates of the relative magnitude of the effects of weather and the epidemic status of the population. Poliomyelitis is now of rather academic interest, but it was very surprising to discover what a very large part of its yearly fluctuations could be ascribed to weather alone.

2.2. *Nonmeteorological seasonal fluctuations.* It is worth emphasizing that seasonal fluctuations may have nothing to do with the weather at all. Several people including, I think, Professor Bartlett have suggested that the seasonal element in measles represents more the progress of the school year than any true meteorological effect. This is almost certainly true of Sonne dysentery, a recurrent minor epidemic disease in English schools with a very well marked seasonal peak in March or April. What appears to happen in these cases is that the disease slowly gathers momentum after the schools assemble in September, reaches a peak in the spring and disappears during the summer holiday dispersal. Many if not most adults and very young children with Sonne dysentery acquire it as a secondary infection from a school child in the same household. A covariance analysis on the lines I have indicated above shows no correlation whatever with any meteorological variables for this disease.

2.3. *Surveillance and forecasting.* The final topic which I would like to discuss is the problem of surveillance and short term forecasting on a routine basis. The larger scale approach that I have considered above needs to be supplemented for practical reasons by some method of detecting as rapidly as possible an untoward increase in the incidence of a disease or group of diseases. Even a week or two of advance notice may be useful, from an administrative point of view, in giving a chance to provide, for example, extra beds for emergency admissions.

Another field in which an early warning system is being increasingly demanded is drug monitoring. In a good many places now there is continuous surveillance of the incidence of congenital abnormalities with a view to detect-

ing as early as possible an increase such as that caused by Thalidomide. Besides this, there is a general tendency, arising from the large scale introduction of new remedies of all kinds, and the realization that animal testing is not a sufficient guide to human reactions, to introduce monitoring schemes covering all new drugs.

The problem as a whole has strong resemblance to the control of quality in manufactured products but has, I think, certain difficulties not encountered quite so acutely in the factory. The underlying "production" process is not at all closely controlled and is afflicted by errors which are usually large and serially correlated. It may also be difficult to find suitable base lines in previous years from which to chart any unfavorable departures. It is likely that no comprehensive statistical solution to these problems is possible and that interpretation of the data will always be partly a matter of experience and judgment.

My own approach to this problem has been quite empirical. Various methods have been suggested in recent years for short term forecasting, for example, in Brown [2] and in various papers by Box and others. In addition to these methods, others have been developed by economists for extracting underlying trends from data affected by random errors and irrelevant seasonal variations. A well known and much used technique of this kind is that developed and programmed for computers by Julius Shiskin [4].

I have tried out these techniques on a number of the common infectious diseases and also on the emergency admissions to hospital in London during the winter months. My experience may have been unfortunate, but I have not found the moving average methods with continuous adjustment of the parameters of much use. They do not seem to be able to adjust themselves quickly enough to the type of change encountered. The Shiskin method was not designed for the study of epidemic disease but is a useful method for producing smoothed graphs freed from seasonal trends. However, the smoothing process used is extremely complicated and its end results difficult to interpret. In my experience it is much more useful in providing a retrospective view than for recognition of future trends.

The method which I have found simplest and most illuminating over the last two years is that of the CuSum chart whose uses in industrial quality control have been described by Woodward and Goldsmith [7]. The essence of this is to plot a cumulative sum of the deviations of the observed incidences from a given base line instead of the individual deviations. This has the effect of smoothing out random errors while rapidly accumulating and showing up systematic departures from the base line. When the errors are not serially correlated and distributed in some known way, tests of significance for departures can be devised. In the case of infectious diseases, the errors are certainly serially correlated and do not follow any very tractable distribution but the method of plotting still seems to be useful in its own right in giving early warning of changes in incidence.

The main difficulty is to provide a suitable base line. Assuming that the main

usefulness of the exercise is to detect excess figures early I have used a set of medium to low incidence years. From these I constructed a series of smoothed figures for each week. The deviations each week from the smoothed values are summed cumulatively and plotted. It is not necessary to have statistical tests of significance, though it would be useful, provided one is only looking for fairly gross changes of incidence.

It is easy to program the CuSum process for computer use and we use this method for the surveillance of congenital malformations. As we have only been doing this for two years we have no historical base line to work from, the method used being based on cumulative sums of deviations of observed values from the national average. Clearly, if some drug such as Thalidomide comes into immediate and universal use throughout the country, its effect would not be noticed, but I think that in fact there would be sufficient variation geographically to give warning if something unusual was going on. The technique as computerized is certainly a very useful one for keeping continuous watch on about 100 types of congenital abnormality in about 150 areas every month of the year but one would be happier if a more solid theoretical basis were available.

2.4. *Necessity for field observations.* I have dealt in this paper with a number of loosely related topics in epidemiology in which some kind of mathematical approach is required to solve specific practical problems. I think I should finish by emphasizing that the development of mathematics in this field is not likely to be successful unless closely related to field observations, and I think a useful outcome of mathematical analysis would be a clear indication of what observations are necessary. If past experience is any guide, these observations will have to be both numerous and expensive. Nor are they likely to be provided by large scale routine notification procedures such as have been the basis of infectious disease statistics in the past.

REFERENCES

[1] N. T. J. BAILEY, "The simulation of stochastic epidemics in two dimensions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1966, Vol. 4, pp. 237–257.

[2] R. G. BROWN, *Smoothing Forecasting and Prediction of Discrete Time Series*, Englewood Cliffs, Prentice-Hall, 1963.

[3] J. M. CRADDOCK, "Analysis of geophysical time series," *J. Roy. Statist. Soc. Ser. A*, Vol. 120 (1957), pp. 387–397.

[4] J. SHISKIN and H. EISENPRESS, "Seasonal adjustments by electronic computer methods," *National Bureau of Economic Research, Technical Paper No. 12*, Princeton, Princeton University Press, 1958.

[5] J. W. TUKEY and R. B. BLACKMAN, *The Measurement of Power Spectra*, New York, Dover, 1958.

[6] M. WISE, "A disease weather analysis," *Int. J. Biometeorology*, Vol. 10 (1966), pp. 77–90.

[7] R. H. WOODWARD and P. L. GOLDSMITH, *Cumulative Sum Techniques*, Ediburgh, Oliver and Boyd, 1964.