# Sepsis in surgical wounds

## Multiple regression analysis applied to records of post-operative hospital sepsis

By O. M. LIDWELL

*Air Hygiene Laboratory, Central Public Health Laboratory, Colindale, London, N.W. 9*

(*Received* 23 *January* 1961)

## INTRODUCTION

The incidence of post-operative wound infection and sepsis in some 3000 patients in twelve hospitals has been the subject of a recent report (Public Health Laboratory Service, 1960). A simple analysis of the records suggested that the risk of sepsis was influenced by a number of factors such as age, duration and type of operation, etc. These factors are, however, highly intercorrelated and it therefore seemed desirable to explore some form of analysis which would give estimates of the independent effect of the several variables. The problem is complicated by the relatively large number of potentially significant factors, by the non-metric character of many of them and by the very unequal distribution of the numbers of operations falling within the various categories. These characteristics are common to much statistical material of medical interest. The availability of electronic high-speed computers now makes possible methods of analysis which were previously quite impracticable because of the excessive labour involved in hand computation. Since these methods have not yet been widely used in problems of this kind, and since there are a number of specific points which arise from their application to this particular set of data, a short description of the method of multiple regression analysis as it may be applied to material of this nature is given, before proceeding to the results obtained when it is applied to the records of post-operative sepsis referred to above.

### *Multiple regression analysis with non-metric variables*

If the result of any trial or observation is dependent on a number of variable conditions then the relationship between the results and the conditions may be expressed in the form

$$Y = f(x_1, x_2, \ldots), \tag{1}$$

where $Y$ is a variable descriptive of the results, excluding error, and $x_1, x_2$, etc., are variables related to the determinative conditions.

If this can be put into the linear form,

$$Y = a + b_1x_1 + b_2x_2 + \ldots, \tag{2}$$

where $a, b_1, b_2$ are constants, then it is easy to derive a series of normal equations from which values of these constants can be derived which will minimize the sum $\Sigma(Y-y)^2$, $y$ being the observed value of the variable $Y$ in a given trial.

The normal equations are:

$$\left.\begin{array}{ll} aN+b_1\Sigma(x_1)+b_2\Sigma(x_2)\ldots & = \Sigma(y), \\ a\Sigma(x_1)+b_1\Sigma(x_1^2)+b_2\Sigma(x_1x_2)\ldots & = \Sigma(x_1y), \\ a\Sigma(x_2)+b_1\Sigma(x_1x_2)+b_2\Sigma(x_2^2)\ldots & = \Sigma(x_2y), \\ a\Sigma(x_3)+b_1\Sigma(x_1x_3)+b_2\Sigma(x_2x_3)\ldots & = \Sigma(x_3y), \\ \text{etc.} & \end{array}\right\} \quad (3)$$

The constant $a$ will be zero if all the variables are expressed as deviations from their mean values. If $a$ is so eliminated then both the first equation in the series and the first term in each of the remaining equations are also eliminated.

Where all the variables involved are continuous variables there are no limitations to the possible functional relationships which may be postulated between the variables, provided that these remain real and finite. For example, analysis into a power series for any variable, $w$, may be carried out by putting $x_1 = w$, $x_2 = w^2$, etc. Interactions between the variables can be included by introducing additional variables which are functions of two or more of the variables.

If some or all of the conditions are non-metric variables, e.g. refer to the presence or absence of particular characters, a similar form of analysis may be carried out by giving to each of the non-metric variables $x_1$,$x_2$, etc., two arbitrarily fixed values, conveniently chosen as 0 and 1, to represent two states. Where more than two states are possible for a given variable then each additional state will need an additional variable in this form.

The formation of the normal equations from data of this kind is extremely simple since all the terms containing the $x$ variables only, e.g. $\Sigma(x_1^2)$, $\Sigma(x_2^2)$, etc., $\Sigma(x_1x_2)$, $\Sigma(x_2x_3)$, etc., are pure frequencies, and can be obtained very readily by sorting, e.g. of edge-punched cards. Thus, $\Sigma(x_1^2) = \Sigma(x_1)$ is the number of trials in which the variable $x_1$ is in the state represented by $x_1 = 1$, $\Sigma(x_2x_3)$ is the number of trials in which the variables defined by $x_2$ and $x_3$ are simultaneously in the states represented by $x_2 = 1$ and $x_3 = 1$ and so on. This simplicity in calculation may make it convenient to handle some continuous variables in this way, i.e. by grouping and then treating each group as a non-metric state. This method may be particularly convenient if the data are only coarsely grouped in the original and if, as is often the case, there is no indication as to the appropriate relationship which should be assumed between the observed quantity and the corresponding variable in the regression equation. Age is an example of a variable to which both considerations often apply.

The constant $a$ in the regression equation is the predicted value of $Y$ when all the conditions are in the state denoted by $x = 0$.

If it is more convenient this constant may be eliminated and a set of reference levels substituted for the single reference state. This can be done by including among the $x$ parameters one group and one only, such that one and only one of the $x$'s within it will be taken as unity whatever the state of the condition to which the group of $x$'s refers.

An example may make the method clearer. Suppose that we are examining the effect of the variables sex and age, and that age is only recorded as under 30, between 30 and 60 or over 60. Then we may take $x_1$ to refer to sex and assign it the value 0 for males and 1 for females. Since age is expressed as one of three states we shall need at least two variables for this factor and may take these as $x_2$ and $x_3$, assigning to $x_2$ the value 1 for ages between 30 and 60 and 0 for any other age, and to $x_3$ the value 1 for ages over 60 and 0 for any other age. Then the constant $a$ in the regression equation will give the predicted value of $Y$ when $x_1, x_2$ and $x_3$ are all zero, i.e. for males under 30. The predicted value of $Y$, for, for example, females over 60, will then be $a+b_1+b_3$. By including a further variable $x_4$ and assigning to it the value 1 for age under 30 and 0 for any other age we obtain a group of $x$'s, $x_2, x_3$ and $x_4$, which form a comprehensive and mutually exclusive set of which one, but never more than one, takes the value 1 whatever the circumstances. The constant $a$ is now eliminated from the regression equation and the predicted values of $Y$ in the cases instanced above are, for males under 30, $b_4$, and for females over 60, $b_1+b_3$.

The result of a trial or observation may itself be, or be expressed as, a non-metric variable, e.g. the occurrence of some event such as death or the appearance of a particular disease. In this case $Y$ can be chosen to represent the probability of the event occurring. The terms containing $y$ in the normal equations are now also simple frequencies, e.g. $\Sigma(x_1 y)$ is the number of times the event occurs among those trials where $x_1$ is reckoned as 1.

The regression equation now takes the form

$$P = Y = a+b_1x_1+b_2x_2\dots, \tag{4}$$

where $P = Y$ is the probability of the event being observed and the $x$'s take the values 0 or 1 according to the states of the respective variables.

This form of the regression equation is the only one for which all the terms in the normal equation are simple frequencies and are unambiguously defined, and it is therefore the easiest to apply in a first examination of the data. It may also adequately represent them.

There are, however, potential absurdities inherent in the expression of a probability in this way. The value of $P$, being the probability of some event, must lie between 0 and 1. There is no such limitation on the values of $P$ as defined by equation (4), and both negative values and values exceeding unity are possible. In some circumstances such 'absurd' values, if they appear, may lie within the range of error of the analysis; in other situations however they may be indicative of a faulty functional relationship.

### *Probability derived by non-additive combination*

Equation (4) is the limiting form, for small values of $p$, of the combination of a set of independent probabilities

$$Q = \alpha q_1^{x_1} q_2^{x_2} \dots, \tag{5}$$

where the $q$'s in accordance with common usage represent the probability of the event failing to occur in association with a particular state of each of the determining conditions. This can be transformed into a linear expression

$$Y = \log Q = \log \alpha + x_1 \log q_1 + x_2 \log q_2 + \dots,$$

which inherently limits $P$ to values less than unity but does not eliminate the possibility of negative values of $P$.

The various factors might, however, combine in a multiplicative rather than an additive manner so that

$$P = ab_1^{x_1}\, b_2^{x_2} \ldots,$$

which is the limiting form, for small values of $P$, of the relationship

$$P = 1-\exp\,(-ab_1^{x_1} b_1^{x_2}) \;\ldots. \tag{6}$$

This can be transformed into a linear regression equation in the form

$$Y = \log\,(-\log \overline{1-P}) = \log\log e + \log a + x_1 \log b_1 + x_2 \log b_2 + \ldots, \tag{7}$$

and the value of $P$ is now effectively limited to the range 0–1.

A difficulty in forming the necessary terms in $\Sigma(x_1 y)$, etc. for the formation of the normal equations derived from equation (7) arises from the fact that the observed value of the probability for a single trial can only take the values 0 or 1 and that log 0 is not finite. This difficulty can often be circumvented in a substantial body of data. Values of $p$ can be obtained from the groups of observations relating to the particular combinations of states of the variables that occur, i.e. $p = (r/n)$ where the event occurs on $r$ occasions out of the $n$ trials recorded under the particular set of conditions. So long as these values of $p$ are not zero or unity no further difficulties arise. When $r$ is zero or equal to $n$ in one or more poorly represented combinations, which together represent only a small fraction of the data, it may be possible to make arbitrary adjustments, e.g. by taking $r$ as $\frac{1}{4}$ or $n-\frac{1}{4}$ in such cases.

A further complication in the use of a regression equation which relates some function of the probability rather than the probability of the event itself to a linear combination of factors is that the normal equations in the form given in equation (3) no longer give the best estimates of the coefficients. For an efficient solution the several groups of observations need to be appropriately weighted before the normal equations are derived. The problem is the same in principle as that discussed by Finney (1947), with reference to probit analysis, and may be solved in a similar manner. A rigorous solution involves however very laborious computation if many variables are involved.

*Testing the regression*

The adequacy of the regression equations can be tested by carrying out a $\chi^2$ test on the individual cells of the classification. As the distribution is binomial the values of $\chi^2$ should be obtained from each cell by applying the formula

$$\chi^2 = \frac{(R-r)^2}{RQ},$$

where $r$ = the observed and $R$ the expected number of occurrences among the $n$ trials within the cell. $Q$ is the expectation that the occurrence will not be found on a single trial under the conditions defining the cell. Values of $\chi^2$ in excess of the expectation of random sampling may be due to errors of observation or recording, the influence of other variables than those included in the regression analysis, interaction between the variables, or other ways in which the assumed regression equation fails to represent the combined effect of the several variables. Visual examination of the observed and expected values in the several cells together with the $\chi^2$ contributions arising from them will reveal any strong interactions and any such which are suspected can be tested by forming suitable multi-way tables and applying the $\chi^2$ test to these. E.g. if two two-way classifications $A$ and $B$ are involved the observed and expected occurrences in the two × two table

| — | $B$ | Not $B$ |
|---|---|---|
| $A$ | — | — |
| Not $A$ | — | — |

are computed.

Where a substantial number of variables are involved and the data are of limited extent it will often happen that some of the cells of the full classification are unrepresented or only poorly represented. In this case such cells may be aggregated into groups with a reasonable expectation, $R \nless 5$, preferably on a random basis, before applying the $\chi^2$ test.

## ANALYSIS OF FACTORS ASSOCIATED WITH POST-OPERATIVE SEPSIS

The data used in this analysis were collected in a survey of post-operation sepsis carried out by the Public Health Laboratory Service and details of procedure, etc., can be found in the report (P.H.L.S. 1960). For each patient records of age and sex and details of the operation performed were available, together with observations on the post-operative state of the wound. A wound was considered septic if it both showed clinical signs of sepsis and yielded pathogenic bacteria on bacteriological examination.

### *Type of operation*

A first analysis was carried out employing a regression equation of the form

$$P = Y = b_1x_1 + b_2x_2 \ldots,$$

and including a selection of those factors which the simple analysis given in the Report suggested were associated with increased incidence of sepsis. The factors included and the classifications of age and operation type employed are given in Table 1. The coefficients found for the regression equation and their standard errors are also given in the table. Since the distributions involved are not normal the error terms are not strictly standard errors but the approximation is sufficiently close to provide an estimate of the significance of the regression coefficients themselves. The estimated risk of sepsis following any particular operation is obtained by adding together the coefficients appropriate to the conditions, e.g. an operation for removal of the appendix performed on a patient aged under 30 taking 45 min. to carry out involving an incision over 6 in. long but needing no drain inserted has an estimated risk of sepsis of

$$0{\cdot}0374 + 0{\cdot}000 + 0{\cdot}0370 + 0{\cdot}0504 + 0{\cdot}000 = 0{\cdot}1248$$

or approximately one in eight.

All of the factors included that might be considered unfavourable are in fact seen to add significantly to the risk of post-operative sepsis although age appears to be a significant factor only in the oldest group.

Table 1. *The effect of the type of operation and other factors*

Coefficients of the regression equation

$$P = Y = b_1x_1+b_2x_2+\dots.$$

| | Variate | No. in class | No. septic | Coefficient ($b$) | Standard error |
|---|---|---|---|---|---|
| 1. | Age 30–59 years | 1542 | 125 | −0·0097 | 0·0147 |
| 2. | Age 60 years and over | 654 | 130 | **0·0778** | 0·0179 |
| 3. | Duration of operation, 31–60 min. | 1089 | 121 | **0·0370** | 0·0139 |
| 4. | Duration of operation, over 60 min. | 704 | 129 | **0·1252** | 0·0186 |
| 5. | Incision over 6 in. long | 804 | 143 | **0·0504** | 0·0147 |
| 6. | Drain inserted | 859 | 163 | **0·0930** | 0·0147 |
| | 7–21. Operation type | | | | |
| 7. | Gall bladder | 213 | 49 | 0·0497 | 0·0272 |
| 8. | Appendix | 484 | 30 | **0·0374** | 0·0144 |
| 9. | Appendix with abscess, etc. | 48 | 21 | **0·3230** | 0·0435 |
| 10. | Gastrectomy | 217 | 21 | **−0·0572** | 0·0273 |
| 11. | Abdominal | 302 | 39 | 0·0008 | 0·0220 |
| 12. | Abdominal with peritonitis, etc. | 47 | 10 | **0·1160** | 0·0442 |
| 13. | Hernia | 383 | 30 | 0·0091 | 0·0199 |
| 14. | Breast | 171 | 26 | 0·0153 | 0·0271 |
| 15. | Varicose veins | 147 | 13 | 0·0367 | 0·0271 |
| 16. | Thyroid | 147 | 6 | **−0·1110** | 0·0295 |
| 17. | Sympathectomy | 35 | 3 | −0·0408 | 0·0501 |
| 18. | Meniscectomy | 62 | 0 | −0·0081 | 0·0375 |
| 19. | Orthopaedic | 194 | 6 | 0·0090 | 0·0230 |
| 20. | Thoracic | 158 | 13 | **−0·1307** | 0·0293 |
| 21. | Any other | 295 | 32 | 0·0078 | 0·0209 |
| | All together | 2903 | 299 | — | — |

Variance absorbed by regression = 61·2.
Residual variance = 237·8.
Numbers 7–21 form an inclusive group.
Those coefficients which exceed twice their standard error are in bold type.

$x_1$ and $x_2$ take the value zero for ages under 30 years, $x_3$ and $x_4$ take the value zero for operations of duration 30 min. or less, $x_5$ takes the value zero for incisions shorter than 6 in. and $x_6$ takes the value zero in the case of those operations where no drain was inserted.

The risk of sepsis varies considerably with the type of operation. A number of operations are associated with negative coefficients of significant magnitude. As a consequence certain combinations of factors are apparently associated with a negative risk of sepsis; for example, the predicted risk of sepsis for a gastrectomy performed on a patient between 30 and 60 years of age taking between 30 and 60 min. to carry out but involving an incision less than 6 in. long and needing no drain is

$$-0{\cdot}0572-0{\cdot}0097+0{\cdot}0370 = -0{\cdot}0299.$$

In fact such groupings usually represent rare or non-occurring combinations of factors. Generally the significance of the negative coefficients is that, for these types of operation, the presence of the various adverse factors does not lead to as high an incidence of post-operative sepsis as might otherwise be expected. Twelve gastrectomy operations of the kind referred to above are actually included in these records; none of these developed post-operative sepsis. As we shall see, however, from the succeeding analysis, these negative coefficients do appear to reflect a real defect in the simple regression equation.

### *Differences between hospitals*

The next group of factors to be examined was the various hospitals in which the operations were performed. This could have been done by carrying out a second regression analysis adding the several hospitals to those variates which appeared significant on the basis of the first. A simpler procedure is to calculate for each of the hospitals the number of cases of post-operative sepsis which would be expected on the basis of the coefficients derived for the various factors from the first analysis and the distribution of those among the patients operated on in each hospital. Neither of these two forms of analysis is entirely proof against false inferences arising out of unexpected strong inter-correlations between any of the omitted variables and those introduced into the second analysis.

Table 2. *Comparison of observed and expected numbers of septic cases according to hospital and sex*

| Hospital | No. of operations | No. septic | % septic | Calculated septic | $\chi^2$ |
|---|---|---|---|---|---|
| 1. R | 782 | 90 | 11·5 | 84·4 | 0·42 |
| 2. W | 689 | 90 | 13·1 | 86·0 | 0·22 |
| 3. PF | 287 | 14 | 4·9 | 30·0 | 8·97 |
| 4. E | 282 | 28 | 9·9 | 30·9 | 0·30 |
| 5. M | 157 | 12 | 7·6 | 11·2 | 0·07 |
| 6. L | 150 | 14 | 9·3 | 12·3 | 0·26 |
| 7. P | 93 | 12 | 12·9 | 8·0 | 2·34 |
| 8. H | 152 | 2 | 1·3 | 2·0 | 1·71 |
| 9. C | 43 | 7 | 16·3 | 3·9 | |
| 10. O | 268 | 30 | 11·2 | 30·3 | 0·00 |
| Total | 2903 | 299 | 10·3 | 299·0 | 14·29 |
| Sex | | | | | |
| Male | 1436 | 158 | 11·0 | 140·3 | 2·50 |
| Female | 1463 | 140 | 9·6 | 157·7 | 2·18 |
| | | | | | 4·68 |

$\Sigma\chi^2 = 14{\cdot}29$ with 8 D.F., $P \simeq 0{\cdot}08$.
$\Sigma\chi^2 = 4{\cdot}68$ with 1 D.F., $P \simeq 0{\cdot}03$.
$\Sigma\chi^2 = 8{\cdot}97$ with 1 D.F., $P \simeq 0{\cdot}002$.

The values of $\chi^2$ have been computed as for a binomial distribution.

The calculated numbers of septic cases have been obtained by using coefficients derived from the analysis whose results are given in Table 3. The preliminary calculations carried out with coefficients derived from the analysis reported in Table 1 led to results insignificantly different from those given above.

Table 2 shows the results obtained from calculations based on the simpler procedure, which was also applied to the division of patients by sex. It is clear that only in one hospital, PF, was there any appreciable discrepancy between the observed and calculated figures. Thus the considerable differences between the ten hospitals in their over-all sepsis rates (column 4) are almost entirely accounted for by the difference in the operations performed in them. In an attempt to see whether any explanation of the more favourable experience in hospital PF could be found in the available data the experience of the patients in the hospital was more closely examined with respect to their age, since it was noticed that there was a more than usually large proportion of children operated on in the hospital. There was, however, no evidence that the experience of this age group was particularly good, low sepsis rates being equally apparent in the adult groups.

It should also be noted that, of the 150 operations performed at hospital L, 146 were thoracic operations. Since the total number of such operations included in the whole data is only 158 it is clearly quite impracticable to attempt to distinguish between this type of operation and this hospital with respect to their effect on the incidence of sepsis.

Although the percentage difference between the calculated and observed numbers of cases of sepsis in male and female patients is not large it is probably significant.

### *Final analysis with the simple regression equation*

As a result of these computations a further analysis was made. The results of this are given in Table 3, which includes all those factors which previous examination had indicated as likely to be significant, including operations involving the gall bladder.

Although Table 3 is concerned with seven fewer variates than Table 1, fourteen in all, the variance absorbed by the regression analysis is actually slightly greater. All the coefficients exceed twice their standard deviation except that for gall-bladder operations where the ratio is only 1·87.

The extent to which the regression scheme of Table 3 is adequate to explain the variations of the original data can be tested by performing a $\chi^2$ test on the data subdivided according to the 192 classifications which this provides. The variates were reduced to 12 for this purpose: no. 13, hospital PF, had to be excluded owing to deficiencies in the classification of the original data and no. 14, sex, was omitted since the coefficient involved was small. Of the 192 classifications fifty-five are not represented and a further 114 have only small expectations, less than five septic cases. These 114 have, therefore, been aggregated at random to produce groups with expectations of 5 or more. As a result of this the $\chi^2$ sum is formed from forty-two cells with 30 degrees of freedom. The sum obtained is 49·9 which, with a probability value of little over 1/100, indicates some excess variance above that due to random sampling. A large part of the excess $\chi^2$ is due to a single cell. Of the 479 unclassified operations carried out, with none of the adverse conditions of age or procedure applying, nine exhibited post-operative sepsis; the regression equation applied to this cell actually predicts a small negative value.

The contribution of this cell to the $\chi^2$ sum is approximately 15·5. The possibility of this kind of failure of the simple form of regression equation has already been referred to.

Table 3. *The association of various factors with the risk of post-operative sepsis*

Coefficients of the regression equation
$P = Y = b_1x_1+b_2x_2+\ldots.$

| Variate | No. in class | No. septic | % septic | Coefficient ($b$) | Standard error |
|---|---|---|---|---|---|
| 1. Age, 60 years and over | 654 | 130 | 19·9 | +0·082 | 0·013 |
| 2. Sex, male | 1438 | 159 | 11·1 | +0·028 | 0·011 |
| 3. Duration of operation, 31–60 min. | 1089 | 121 | 11·1 | +0·028 | 0·013 |
| 4. Duration of operation, over 60 min. | 704 | 129 | 18·3 | +0·109 | 0·018 |
| 5. Incision over 6 in. long | 804 | 143 | 17·8 | +0·056 | 0·014 |
| 6. Drain inserted | 859 | 163 | 19·0 | +0·111 | 0·014 |
| 7. Appendicectomy with abscess or peritonitis in an abdominal operation | 95 | 31 | 32·6 | +0·180 | 0·031 |
| 8–13. Operation type | | | | | |
| 8. Appendix | 532 | 51 | 9·6 | +0·044 | 0·015 |
| 9. Gall bladder | 213 | 49 | 23·0 | +0·038 | 0·024 |
| 10. Any other | 1636 | 159 | 9·7 | −0·006 | 0·012 |
| 11. Gastrectomy | 217 | 21 | 9·7 | −0·072 | 0·026 |
| 12. Thyroid | 147 | 6 | 4·1 | −0·114 | 0·028 |
| 13. Thoracic | 158 | 13 | 8·2 | −0·163 | 0·029 |
| 14. Operation performed at hospital PF | 287 | 14 | 4·9 | −0·076 | 0·019 |
| All together | 2903 | 299 | 10·3 | — | — |

Variance absorbed by regression = 61·9.
Residual variance = 237·1.
Numbers 8–13 form an inclusive group.
$x_1$ takes the value zero for ages of 60 years and under, $x_2$ takes the value zero for females, $x_3$ and $x_4$ take the value zero for operations of duration 30 min. or less, $x_5$ takes the value zero for incisions shorter than 6 in., $x_6$ takes the value zero in the case of those operations where no drain was inserted and $x_7$ takes the value zero for all operations except those involving appendicectomy with an abscess or abdominal operations with peritonitis.

### *Analysis employing non-additive combinations of factors*

Instead of the risks due to the various factors combining additively we might assume the multiplicative form of combination defined by equation (6) above. That this form of combination may be more appropriate in the present instance is suggested by the results obtained on considering each of the potentially adverse factors as equally powerful (operations lasting longer than 60 min. are considered to involve a further adverse factor in addition to that for operations lasting more than 30 min.). The relationship between the risk of sepsis and the number of adverse factors is shown in Fig. 1, and it will be seen that the risk increases more rapidly than the number of adverse factors in a way consonant with the multiplicative form of equation (6) given above.

An unweighted analysis based on the regression equation (7) above was then carried out employing the adjustment for extreme values of $p$ described earlier. The results of this are given in Table 4. The constants evaluated in this way give a minimum $\Sigma(Y-y)^2$ which is not identical with a maximum likelihood or a minimum $\Sigma(R-r)^2$ solution. When the expected numbers of septic cases $\Sigma(R)$ are calculated for the individual cells in order to carry out a $\chi^2$ test for goodness of fit $\Sigma(R)$ does not equal $\Sigma(r)$ either over-all or in the single factor groupings. In spite of this, however, the $\chi^2$ test carried out in the same way as for the additive regression equation gives a sum of 35·1 for thirty-eight cells with 26 degrees of freedom which with a probability a little greater than 0·1 is a considerable improvement in fit over that obtained with the assumption of an additive combination of factors.

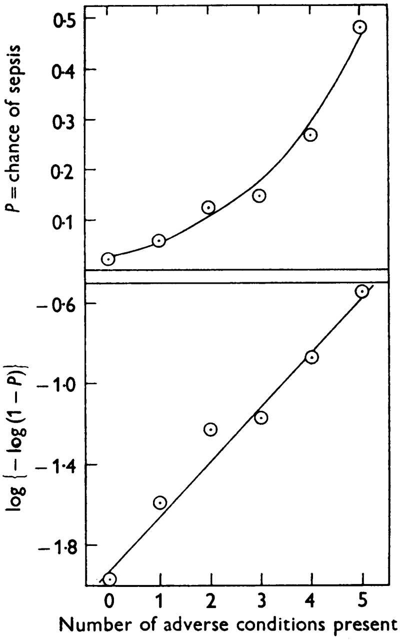

Fig. 1. The risk of sepsis as a function of the number of adverse factors present at operation. The factors included are: age over 60, duration of operation, incision over 6 in. and insertion of a drain.

An analysis was also performed using weights equal to the observed responses, i.e. the values of $r$, with the same adjustment for extreme values. The $\chi^2$ test, however, when applied to the results of these calculations, resulted in a somewhat greater $\chi^2$ sum and a consequently lower value of the probability. The better result obtained in this instance from the unweighted analysis is likely to be fortuitous, both solutions are inefficient, but it does show that a reasonably good approximation can be obtained in this way. Using the results of Table 4 as a basis, an attempt was made to improve the fit between the observed numbers of septic cases in the various classifications and those calculated on the basis of equation (6) by making arbitrary adjustments to values of the constants. Probability values up to 0·2 were reached. This represents a very fair fit, especially since two factors—sex and operation performed in hospital PF—known to be significant, have been omitted. There does not seem, therefore, to be any reason in this instance to carry out any more rigorous solution of the coefficients of the regression equation.

Table 4. *The association of various factors with the risk of post-operative sepsis*

Coefficients of the regression equation

$$Y = \log\{-\log(1-P)\} = x_1 \log b_1 + x_2 \log b_2 \ldots$$

| | Variate | Coefficient (*b*) |
|---|---|---|
| 1. | Age, 60 years and over | 2·26 |
| 2. | Duration of operation, 31–60 min. | 2·00 |
| 3. | Duration of operation, over 60 min. | 3·68 |
| 4. | Incision over 6 in. long | 1·79 |
| 5. | Drain inserted | 2·05 |
| 6. | Appendicectomy with abscess or peritonitis in an abdominal operation | 2·71 |
| | 7–12. Operation type | |
| 7. | Appendix | 0·0351 |
| 8. | Gall bladder | 0·0333 |
| 9. | Any other | 0·0228 |
| 10. | Gastrectomy | 0·0116 |
| 11. | Thyroid | 0·0066 |
| 12. | Thoracic | 0·0086 |

The factors numbered 7–12 are an inclusive set and hence form a set of reference levels to which the multiplicative effects of the other factors are applied. The coefficients $b_7$–$b_{12}$ are approximately equal to the risk of post-operative sepsis associated with each type of operation in the absence of any of the factors 1–5 and the effect of these factors is approximately given by multiplying the risk for the operation type by the corresponding coefficients.

$x_1$ takes the value zero, equivalent to $b_1$ equal to unity, for ages of 60 years and under, $x_2$ and $x_3$ take the value zero for operations of duration 30 min. or less, $x_4$ takes the value zero for incisions shorter than 6 in., $x_5$ takes the value zero in the case of those operations where no drain was inserted and $x_6$ takes the value zero for all operations except those involving appendicectomy with an abscess or abdominal operations with peritonitis.

## DISCUSSION

The more detailed analysis described in this paper has, for the most part, confirmed the results obtained by the simpler form of evaluation presented in the original report. Thus age over 60 years, duration of operation exceeding 30 min., incision over 6 in. long and the insertion of a drain are all now found to be independently associated with a significantly increased risk of sepsis. There are, however, a number of points where the regression analysis leads to conclusions which differ from the earlier indications. The small and non-significant difference in the over-all sepsis rate of the two sexes is increased to a significant level when allowance is made for the simultaneous effect of other factors, so that males appear to be appreciably more liable to develop sepsis than females. The considerable differences between the post-operative rates in the various hospitals are seen to be almost entirely accounted for by the differences between them in respect of the characteristics of the patients and the types of operation performed on them reported in this survey. Only one hospital appears to differ significantly from the others when these factors are allowed for. The relative liability to sepsis associated with the different types of operations in themselves is also considerably modified by the regression analysis. Thus appendicectomy is found to carry the highest risk of all operation types although the over-all sepsis rate for this type of operation is below the average, it being less often associated with other adverse factors.

The method of multiple-regression analysis employed in evaluating the data discussed here is very easily and expeditiously carried out if the services of an electronic computer are available. This is still largely true even if the more complex form of regression equation, in which the predicted probability is derived by a multiplicative combination of factors, is substituted for the simpler additive form, so long as an unweighted solution of the equations is adequate. The multiplicative form employed has the advantage that the predicted probabilities are inherently confined within the natural range of 0–1; it was also found to give a significantly better description of the observed data.

## SUMMARY

The method of multi-regression analysis as applied to data which are classified in a non-metric manner is discussed, and a set of data relating the risk of post-operation sepsis to various factors is analysed in this way.

Age over 60, male sex, long duration of operation, incision over 6 in. long and the insertion of a drain in the wound are all found to be associated with increased risk of post-operative sepsis. Operations for appendicectomy and those on the gall bladder have a relatively high basic risk of developing sepsis while operations for gastrectomy or those on the thyroid, and probably thoracic operations, have a low basic risk. The substantial differences in the gross risk of post-operative sepsis associated with the operations performed in the different hospitals included in the analysis are shown to be almost entirely accounted for by differences in the operations performed in them.

Our thanks are due to the department of statistics of the Rothamsted Experimental Station, and in particular to Mr J. H. A. Dunwoody and the computing staff who carried out the calculations on the electronic computer.

## REFERENCES

Finney, D. J. (1947). *Probit Analysis*. Cambridge University Press.
Public Health Laboratory Service (1960). *Lancet*, ii, 659.