THE RELATION BETWEEN THE CANCER AND DIABETES DEATH-RATES.

By M. GREENWOOD, Junr.
and FRANCES WOOD.

(From the Statistical Department of the Lister Institute of Preventive Medicine.)

Introduction.

Early in 1912 Professor E. C. C. Baly, F.R.S., acting on behalf of Mr Jessup, brought one of us a collection of Swiss statistics and requested that it should be analysed, with the object of measuring the relation between the death-rates of cancer and diabetes. In the course of this work we have been led to consider many problems of theoretical and practical interest, bearing partly on the method of analysis and partly on the results yielded; it would not be possible to deal with all these in a single paper and the present memoir is devoted to a study of the problem first suggested, viz. the relation between the cancer and diabetes death-rates. Since, however, this may be the first paper of a series, it will be convenient to deal at some length with the nature of the investigation and the motives which led us to undertake it.

In 1909 Dr G. D. Maynard published under the title of "A Statistical Study of Cancer Death Rates" (*Biom.* 1909–10, vii. pp. 278–304) an important contribution to the subject now under discussion.

In the first place Maynard set himself to determine whether the incidence of cancer varied with meteorological conditions. His material consisted of cities in the United States of America, each having over 100,000 inhabitants and he correlated the corrected cancer rate with sunshine in hours, mean temperature, rainfall and other meteorological conditions. In no case was a definitely significant result obtained.

It then occurred to him to correlate the cancer rate with that of

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

some other diseases and he was particularly led to consider diabetes in this connection, because (1) both diseases have very much the same age distributions, (2) both are on the increase, (3) the aetiology of both diseases is obscure.

The method he adopted was twofold. In the first place he correlated the absolute number of deaths from cancer (c) with the absolute number of deaths from diabetes (d), then correlated each with the absolute population (p) and finally worked out the partial correlation of cancer deaths with diabetes deaths for population "constant," *i.e.* ${}_pr_{cd}$. In the second place the absolute numbers of deaths from each disease were corrected for age distribution by the use of the well-known correlation factors (for description see supplement to 65th annual report of the Registrar General) and the previous process repeated with these corrected totals.

These methods were employed in the case of 15 states and 40 cities. The results were that in every case a large and significant positive correlation was found. Thus, using uncorrected figures, the values were, 6896 ± ·0559, ·9088 ± ·0303 for cities and states respectively; and using corrected figures, 7325 ± ·0494 and ·8258 ± ·0554.

These very high values either (1) might indicate a close relationship between the two diseases in consequence of some common physiological factor; (2) might be due to some error of method or material; or (3) might be dependent upon some indirect cause of association, *e.g.* the influence of common occupation.

The first possibility which suggested itself was that some element of so-called spurious correlation had been introduced by the process of correcting for age distribution, which involved the application of certain common factors to both of the variables. This possibility has been made the object of a special inquiry by Pearson and Lee (*J. Roy. Stat. Soc.* 1910, LXXIII. 534) in which an elaborate method of neutralising this source of error is described and illustrated. The correction was applied to Maynard's data and it was found that no appreciable change was made in the value of the correlation. Pearson accordingly concluded that Maynard's results could not be accounted for in this way. We have ourselves further considered this point and have applied several tests to Maynard's results. Our work entirely confirms the view of Pearson and we have no doubt that whatever may be the cause of Maynard's results, methodological errors cannot have been significant sources of correlation. The nature of the data is a much more disputable subject.

The first question is as to whether the returns might not be untrustworthy and not real measures of the prevalence of the two diseases. Maynard investigated this point in the following way. In the first place he thought the registration of deaths might be very imperfect. This had been the subject of a special inquiry by the United States Census department in 1900. Enumerated deaths were compared with registered deaths and the result showed that, so far as the cities investigated by Maynard are concerned, the error never exceeded 9·5 %; in only one city was the error greater than 6·5 %, while the mean was 3·5 %, with a coefficient of variation of 2·225. So that it would appear that the numbers of deaths which escape registration are not large enough to be a serious source of error. This does not of course enable us to form an opinion as to the accuracy of the diagnoses recorded on the certificates of death. As a check on this Maynard correlated the cancer and diabetes rates with deaths from ill-defined or unknown causes and found that the correlation was very small in both cases. In reply to the objection that in towns well equipped with competent medical men, the two diseases would figure less in the death returns because of operative treatment in the case of cancer and dietetic treatment in the case of diabetes and that the converse would hold in towns where medical skill was inferior or the population too ignorant to follow advice, Maynard urges that these cities do not probably differ greatly in such respects and that in neither disease is a cure effected, even by the best treatment, except in rare instances.

The most obvious cause of indirect association, other than those already mentioned, would be that certain occupations predispose to the development of both diseases. To test this Maynard correlated occupations with diabetes and with cancer rates. He found that the coefficient differed not only absolutely, but also in sign, *i.e.* that the occupations with a high cancer rate tended to exhibit a low diabetes and conversely. It did not therefore seem possible to explain the results in this way.

Having apparently exhausted the indirect or erroneous causes of correlation, Maynard offered a provisional explanation of his results. He writes: "Only one cause, it seems to me, will fit the facts as we know them, viz. the pressure of modern civilization and the strain of modern competition or some factor closely associated with these."

He suggests that the increasing tension of life owing to competition and nervous strain might account for the facts and points out that there is a high correlation between either rate and the prevalence of insanity

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

in the states used and that there is a similar correlation with the number of suicides in both states and cities.

Another point is whether the correlation might not be due to both rates increasing with the general unhealthiness of the district. This does not seem to be the case, because (1) the correlation between cancer and nephritis is very much smaller than between cancer and diabetes, and (2) Pearson found from Maynard's data that the correlation between cancer and deaths from all causes other than diabetes and cancer was markedly less than that between cancer and diabetes.

As will be seen, Maynard was thoroughly on his guard against material fallacies and both he himself and also Pearson have carefully and, we think, successfully guarded against the risk of error due to improper analytical treatment; the one point which we think it advisable to subject to further scrutiny is the question of diagnostic accuracy. The authorities of the Imperial Cancer Research Fund have referred to the curious difference between the recorded death-rates from cancer in Ireland and in the rest of the United Kingdom (*Scientific Reports on the Investigations of the I.C.R.F.* No. 2, Part 1; *The Statistical Investigation of Cancer,* by Drs Bashford and Murray, pp. 40–3). They point out that the death-rate from cancer in Ireland is very much smaller than in the rest of the United Kingdom. It is then noted that autopsies are performed in the Irish hospitals much less frequently than in England. Since many cases would not be diagnosed but for the post-mortem findings, it follows that the fewer the autopsies the fewer will be the proportional numbers of cancer cases diagnosed. The authors write:

"A general review of the facts elicited on the diagnosis of cancer in London and Irish Hospitals respectively reveals the magnitude of the differences which may exist between the recorded frequency of cancer and its absolute incidence. Procedures which in London and England generally make the recorded incidence approximate nearly to the true incidence are inoperative in Ireland. The discrepancy between the recorded frequency of cancer in the hospitals of London and Ireland can be explained solely by the disadvantages under which the search for cancer is carried out in the latter country."

The authors publish some hospital statistics which demonstrate that many cases of malignant disease will be overlooked if a post-mortem examination does not form part of the inquiry. The data suggest (if we have correctly read the table on p. 14) that possibly the number recognised on clinical examination would be increased by

21 % as a result of simple post-mortem examination or by 35 % if the latter be supplemented with microscopic examination (312 became 377 and 338 became 456 respectively). An increase of one-third is not sufficient to bring the Irish rate (taking the proportion revealed among in-patients of hospitals) up to that of England. Simply on the evidence actually adduced, the writers' statement is, perhaps, unproven; this is by no means to say that it is false and we have referred to it here because it points to an important *possible* source of error in Maynard's data. Maynard's investigation of the influence of bad diagnosis, although evidently scientific and sufficient so far as it goes, does not, we think, go far enough. The low correlation between the cancer rate and the rate of cases ill-defined or not defined does not guarantee the accuracy of the diagnosis. Such a correlation might only indicate, if we may to make the point clear be allowed to state it rather flippantly, that the cancer rate was not affected by the conscious ignorance of the medical practitioner. A high proportion of undiagnosed causes of death may not be so much a measure of ignorance *per se* as of conscious ignorance. It is evidently of importance to determine whether there really is any appreciable difference between the amount of medical skill available in the different cities used by Maynard. How this might be expected to operate can be readily seen. Suppose we had only two cities in any district in one of which there was an active and scientific medical school and in the other no medical school or an inferior one. Not only will patients tend to flow towards the former for their own sakes, but in addition there will be attempts on the part of the authorities of the scientific school to attract cases for the purpose of study and demonstration. Both cancer and diabetes are diseases which an active school might desire to have in its wards. It is no doubt perfectly true, as Maynard remarks, that little or nothing can be done for diabetes but the disease is one of peculiar interest to chemical pathologists to the very end of life, owing to the phenomena associated with diabetic coma. The interest and importance of every phase in the life history of a person suffering from cancer are of course obvious. All these considerations suggest that an apparent correlation between the diabetes and cancer rates might very well be a secondary consequence of an association between both and the presence of efficient medical schools and we therefore specially investigated this point.

We first attempted to sort out the cities on the basis of the laws governing medical practice in the various states of the Union. These laws have been epitomised in a publication issued by the English General Council for Medical Education. A careful study of them

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

convinced us that we could not with their aid determine which cities were better or worse supplied with scientifically trained doctors. We then considered the state of the medical schools in the various cities. Here we had better material to use, in the shape of Flexner's searching inquiry into the condition of medical education in the United States. With the help of Flexner's descriptions we were able to classify the cities used by Maynard into those containing undoubtedly efficient medical schools and those the medical schools of which were either inefficient or non-existent. We thus had data for determining the correlation between the efficiency of the medical school and the rate of cancer or diabetes. The correlation could, however, only be determined by Pearson's two row method which assumes, (1) that the regression is linear, and (2) that the variable given only in two categories is approximately normal in distribution. Both these assumptions being of doubtful validity in the present case, the resulting coefficient can hardly be compared with others calculated from product moments in the ordinary way. The results appear in Table I.

Since evidently the correlation between goodness of medical school and height of the cancer rate might be indirect and due to the correlation of both variables with the size of the city, we desired to find the partial correlation of medical schools with cancer rate keeping population constant. But this involved the introduction of a coefficient calculated on a different basis, viz. by direct product moments, and introduced a risk of the coefficients themselves being, as it were, heterogeneous; however, the result, for what it is worth, is that the correlation between presence of a good medical school and the cancer rate for constant population is $\cdot 2349 \pm \cdot 1008$ (see Table I).

We think this finding is evidence that there is some correlation between the efficiency of the medical school and the cancer rate.

In order to get some collateral check we have turned to English cities and Metropolitan boroughs with populations over 200,000, and have compared the rates in cities with medical schools with those in cities without medical schools.

Both the English and American results agree in showing that there is a small but distinct correlation between the presence of a medical school and the cancer death rate, a fact which, so far as it goes, substantiates the view of the Imperial Cancer Research workers that the actual rate of cancer is influenced by the frequency and efficiency of post-mortem examinations. But the correlations we have determined decisively and we think finally negative the idea that Maynard's high

TABLE I.

Correlation between presence of a good medical school in a city and the mortality from cancer and from diabetes respectively. Males and Females.

(1) 40 *American Cities*[1].

Mean death-rate[2] from cancer for all cities	= 732·625
Mean death-rate from cancer for 20 cities with good medical schools	= 771·500
Standard deviation of the cancer death-rates for all cities ...	= 128·671
Number of cities with reputable medical schools	= 20
Correlation between death-rate from cancer and the presence of a good medical school	= +·3787[3]
Mean death-rate from diabetes for all cities	= 105·850
Mean death-rate from diabetes for 20 cities with good medical schools	= 108·950
Standard deviation of the diabetes death-rates for all cities ...	= 32·151
Correlation between death-rate from diabetes and the presence of a good medical school	= +·1208[3]
Mean population for all cities...	= 374·00
Mean population for 20 cities with good medical schools ...	= 574·45
Standard deviation of the population for all cities	= 449·24
Correlation between population and the presence of a good medical school	= +·5592[3]
Correlation between death-rate from cancer and population ...	= +·3511[4]
Correlation between death-rate from cancer and the presence of a good medical school keeping the population constant ...	= +·2349[5]

(2) 22 *English Cities and Boroughs having populations over* 200,000[6].

Mean death-rate[2] from cancer for all cities	= 995·36
Mean death-rate from cancer for 10 cities with medical schools ...	= 1023·00
Standard deviation of the cancer death-rates for all cities ...	= 79·62
Correlation between death-rate from cancer and the presence of a medical school	= +·3980[3]

[1] The corrected death-rates from cancer and diabetes used in these calculations were obtained from a paper by Dr Maynard of Pretoria entitled "A Statistical Study in Cancer Death-Rates" (*Biom.* VII. p. 276). Mr Abraham Flexner's Report on "Medical Education in the United States and Canada" was used for determining which cities contain reputable medical schools.

[2] Corrected death-rate per 1,000,000 living.

[3] Calculated by Professor Pearson's method for cases when one variable is given quantitatively and the other in alternate categories (*Biom.* VII. p. 96).

[4] Calculated by the ordinary product moment method.

[5] Using the ordinary formula $_z r_{xy} = \dfrac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$.

[6] The deaths from cancer for the different cities were kindly supplied by the Registrar General of England and Wales. The actual figures are given in Table XI.

correlations can have been seriously affected in this way. We have found that there is no appreciable correlation between the presence of medical schools and the diabetes death-rate (Table I); further, and this seems to amount to a complete demonstration of the point, if we select from Maynard's data those cities *with* good medical schools and correlate the diabetes and cancer rates for those cities alone, the result is almost exactly the same as yielded by the whole of the cities taken together (see Table II).

TABLE II.

Correlation between the corrected death-rates from cancer and from diabetes in 20 American cities all possessing good medical schools. Males and Females.

Mean death-rate from cancer = 771·50	Standard deviation of cancer death-rates = 132·97
Mean death-rate from diabetes = 108·95	Standard deviation of diabetes death-rates = 34·99

Correlation between death-rate from cancer and from diabetes in 20 American cities possessing good medical schools	= + ·6392 ± ·0892
Correlation between death-rate from cancer and from diabetes for all the 40 cities[1]	= + ·6802 ± ·0573

[1] Calculated from data given in Dr Maynard's paper.

We now turn to an extremely important matter, viz. the influence on the observed correlations of racial mixture. The United States of America differ from all other civilized countries in the extent to which the inhabitants are foreign born and in the numbers of different nationalities represented in their population. A simple illustration is afforded by the fact that in 37 cities, which we have specially investigated, the mean proportion of the inhabitants who were children of foreign born parents was 57 %. We have no strictly comparable data for other countries, but it seems in the highest degree improbable that a similar state of affairs prevails in any of those we have studied. Now it is readily conceivable that these conditions might exert an effect upon the correlation between diabetes and cancer even if the relation between the diseases themselves were not direct. Thus, considering two races, A and B, let us suppose that in the former the liability to develop cancer were high and in the latter the liability to die of diabetes were great. Then, if immigrants from A and B tended to pass into the same cities in about equal numbers a high correlation between the death-rates from the two diseases might be produced. The same remark would of course apply if in any given nationality both rates tended to be high; this would

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

happen if the two diseases were dependent on or related to the stage of culture attained by a nation.

A solution of this problem is of the first importance but materials for obtaining it do not seem to exist. We know the numbers of foreigners derived from different nations in each city, but we do not know either their separate age distributions or the death-rates from the two diseases which should be applied to them. It would be futile to use the rates obtaining in the countries of origin, since the figures are probably not valid for international comparisons.

We have only been able to approach this problem indirectly and our results are inconclusive. We adopted the following method:

The different nationalities were divided into 13 groups and we then determined for each city what Pearson has defined as a coefficient of intra-racial heterogeneity (*Biom.* 1907, v. 198). When this coefficient is high it means that the corresponding city differs greatly in racial composition from the general constitution of the whole population in the cities studied and when it is low the inference is the converse. We then ranged the cities in the order of their coefficients of heterogeneity and divided them into two equal groups—the one containing the more and the other the less divergent cities. The correlation between the corrected death-rates from cancer and diabetes was then calculated separately for each group and found not to differ significantly from that given by all the cities taken together (Table III). ($\cdot 6044 \pm \cdot 1009$ (less heterogeneous), $\cdot 7103 \pm \cdot 0788$ (more heterogeneous), $\cdot 6769 \pm \cdot 0542$ (all cities together).)

There was a slight diminution in passing from the more to the less heterogeneous but not a significant one. We repeated the process with a smaller group of 11 cities having smaller coefficients of heterogeneity, but once more no significant difference was revealed (·6115).

In interpreting these results it should be remembered that the word heterogeneity is used in a rather ambiguous sense. Because a city has a small coefficient of heterogeneity it does not mean that its population is homogeneous; far from it; but merely that it does not differ markedly from the racial distribution characterising the whole "population" of cities. Had we found a significant result we should have been entitled to say that intra-national heterogeneity (*i.e.* relative heterogeneity) is a factor of importance—a failure to discover this by the present method is not, however, decisive[1]. To settle the matter, details of the

[1] The point is rather *how* the heterogeneity is produced than its magnitude as revealed by the above method.

TABLE III. United States of America.

A. *Correlation between corrected death-rates from cancer and from diabetes in 36 American cities allowing for the racial heterogeneity of the different cities. Males and Females.*

(1) *Correlation for 18 cities with coefficients of racial heterogeneity less than ·1*[1].

Mean death-rate from cancer = 673·278	S.D. of cancer death-rates = 90·541
Mean death-rate from diabetes = 96·944	S.D. of diabetes death-rates = 27·804

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6044 \pm \cdot 1009.$$

(2) *Correlation for 11 cities with coefficients of racial heterogeneity less than ·02.*

Mean death-rate from cancer = 683·727	S.D. of cancer death-rates = 82·936
Mean death-rate from diabetes = 104·000	S.D. of diabetes death-rates = 28·800

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6115. \quad \text{S.D. } \cdot 1980^{2}.$$

(3) *Correlation for 18 cities with coefficients of racial heterogeneity greater than ·1.*

Mean death-rate from cancer = 784·389	S.D. of cancer death-rates = 133·365
Mean death-rate from diabetes = 109·500	S.D. of diabetes death-rates = 33·735

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 7103 \pm \cdot 0788.$$

(4) *Correlation for the 36 cities*[3] *taken together.*

Mean death-rate from cancer = 728·833	S.D. of cancer death-rates = 126·800
Mean death-rate from diabetes = 103·222	S.D. of diabetes death-rates = 31·543

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6769 \pm \cdot 0542.$$

B. *Correlation between corrected death-rates from cancer and from diabetes for a constant proportion of foreigners in 37 American cities*[3].

Mean death-rate from cancer = 731·649	S.D. of cancer death-rates = 126·211
Mean death-rate from diabetes = 103·378	S.D. of diabetes death-rates = 31·049
Mean proportion of foreigners = 57·432	S.D. of proportion of foreigners = 19·680

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6763 \pm \cdot 0610.$$

$$r_{\frac{c}{p}f} = +\cdot 4003 \pm \cdot 0944.$$

$$r_{\frac{d}{p}f} = +\cdot 3598 \pm \cdot 0979.$$

$$\therefore {}_{f}r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6225 \pm \cdot 0689.$$

[1] For the actual values of the coefficients see Table IV.

[2] See footnote (3) to Table XIII (p. 111).

[3] The proportion of coloured persons could not be obtained for 4 cities, so that coefficients of racial heterogeneity could only be determined for 36 cities. The proportion of foreigners was obtained for 37 cities. The data for the calculations were obtained from the 12th Census of the United States of America, 1901–1902.

C. *Correlation between corrected death-rate from cancer and from diabetes for a constant proportion of Colour and Irish in 36 American cities.*

Mean proportion of Coloured[1] = 7·6183	S.D. of the Coloured = 11·2600
Mean proportion of Irish = 10·3792	S.D. of the Irish = 6·7425

$$r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6769 \pm \cdot 0609. \qquad r_{\frac{d}{p}Cl} = -\cdot 2592 \pm \cdot 1049.$$

$$r_{\frac{c}{p}Cl} = -\cdot 1618 \pm \cdot 1095. \qquad r_{\frac{d}{p}I} = +\cdot 3454 \pm \cdot 0990.$$

$$r_{\frac{c}{p}I} = +\cdot 2473 \pm \cdot 1055. \qquad r_{Cl\,I} = -\cdot 4344 \pm \cdot 0912.$$

$$\therefore {}_{Cl\,I}r_{\frac{c}{p}\frac{d}{p}} = +\cdot 6491 \pm \cdot 0650.$$

[1] See A (4) for means and S.D. of the cancer and diabetes death-rates.

TABLE IV. UNITED STATES OF AMERICA.

Corrected cancer and diabetes death-rates, coefficients of racial heterogeneity and condition of medical education in 40 American cities[1].

Cities with at least one reputable medical school. (Flexner)				Cities with either poor medical schools or none. (Flexner)			
City	Cancer death-rate	Diabetes death-rate	Coefficient[2] of racial hetero-geneity	City	Cancer death-rate	Diabetes death-rate	Coefficient of racial hetero-geneity
Queen's	565	106	—[3]	St Joseph	446	58	·064
Indianapolis	596	95	·072	Memphis	497	55	·174
Omaha	659	73	·057	Alleghany[4]	537	59	·046
Pittsburg	686	55	·046	Scranton[4]	561	113	·057
Washington	703	99	·184	Jersey City[4]	627	94	·059
Cleveland	703	66	·153	Louisville	648	78	·095
Columbus	702	85	·063	Toledo	653	115	·057
St Louis	722	74	·090	Kansas City	652	70	·070
Philadelphia	723	98	·127	Denver	712	125	·050
Brooklyn	725	128	—[3]	Paterson[4]	719	100	·108
Syracuse	735	111	·036	Minneapolis	721	89	·199
New Haven	774	148	·061	St Paul[4]	724	87	·095
Baltimore	793	101	·130	Fall River[4]	753	131	·185
Detroit	807	122	·116	Newark[4]	769	119	·051
New Orleans	812	65	·168	Worcester[4]	756	132	·088
Buffalo	820	163	·117	Rochester[4]	808	132	·053
Chicago	840	100	·216	Cincinnati	811	91	·100
Boston	937	151	·179	Richmond	816	142	·126
Manhattan	944	175	—[3]	Providence[4]	832	156	·242
San Francisco	1184	164	·121	Milwaukee	833	109	—[3]

[1] Classified by means of Flexner's report.

[2] For method of calculation see *Biom.* v. p. 198. Data for this calculation were obtained from 12th Census of the United States of America, 1901-2.

[3] Coefficient in this case could not be calculated as the necessary data could not be obtained for the particular unit used for the calculation of the death-rates.

[4] Had no medical school.

age distribution and rates of mortality from cancer and diabetes for each of the nationalities represented in the cities would seem essential. We could then correct for both age and nationality. We also measured the effect of keeping the proportions of certain largely represented nationalities constant. Thus coloured persons and Irish are largely represented in many of the cities.

We worked out the correlation between the two rates—keeping the proportion of Irish and coloured persons constant. The coefficient is $\cdot 6491 \pm \cdot 0650$ (see Table III). No real change has been effected. It is vexatious not to be able to arrive at more definite results upon this point, since it is hard to escape from the impression that this factor of racial heterogeneity may have been extremely important in leading to the results, but with the data at our disposal it seems quite impossible to arrive at scientifically valid conclusions. The opinion may perhaps be ventured that, for the purpose in hand, the racial heterogeneity of the American cities introduces a disturbing influence which renders the data less suitable than some of those we have used.

In concluding our review of Maynard's pioneer investigation we desire to pay a tribute to its value and importance. Maynard was the first medical writer to apply exact statistical methods to the elucidation of the cancer problem and his paper might well serve as a model for those desiring to come to close quarters with this important branch of study. We shall now turn to our own investigations.

Ideal data for the study of this problem would conform to the following standards:

(1) There should be a large number of districts or towns for each of which the cancer and diabetes rates are known.

(2) The population in each district should be so large or the record should extend over so long a period of time that the actually recorded deaths may reasonably be taken as accurate measures of the prevalence of the diseases studied.

(3) It is desirable that all the populations should be large enough to fulfil the requirements of (2), but that they should not vary enormously from district to district.

(4) It is necessary that the state of medical knowledge and the organization and control of vital statistics should be such that we can regard the records as reasonably trustworthy.

(5) It is desirable that the districts used should not exhibit marked racial and economic heterogeneity.

The importance of all these considerations, except (3), is too obvious to need detailed justification. With regard to (3) the point is a statistical one. When the populations differ very greatly in size the standard deviation becomes large relatively to the mean and this introduces analytical difficulties in comparing correlations based upon absolute numbers with coefficients based on proportional frequencies (*vide infra*). We shall now consider how far our material conforms to the above standards.

We have examined most of the published vital statistics of civilized countries and have used the following:

(1) Vital Statistics of the Swiss Confederation;

(2) Vital Statistics of Italy;

(3) Vital Statistics of England and Wales.

Contrary to expectation, neither the German nor French vital statistics are tabulated in a manner suitable for the purpose of the present research.

Switzerland.

The advantages of the Swiss material are numerous. In the first place, there are reasons to believe (we have received assurances on this point from more than one authority on the subject) that these data are from the medical point of view to be compared favourably with those of any other country. It is probable that the recorded cancer incidence differs less from the real incidence than elsewhere. In the second place, thanks to the co-operation of Mr Jessup and the courtesy of the officials of the Swiss Statistical Bureau, we have far more detailed information respecting these data than in any other case. For these reasons we should be inclined to accord to Switzerland a leading place in order of importance.

There are, however, some grave disadvantages to be set against the merits detailed. These mainly depend upon the fact that Switzerland is a very small country; we can tabulate the figures for only a small number of districts, and in addition the absolute populations of these districts are in several cases very small, only a few thousands where we should like to see hundreds of thousands. In the case of so common a disease as cancer this may not be of much importance; we might suppose that the recorded rates approximate fairly closely to the real rates and that the effects of random sampling have not been of moment although in the light of further results we are by no means confident of this. Certainly the statement is not true in the case of a rather

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

uncommon disease like diabetes. In that case the recorded rate depends in a few instances upon less than ten recorded deaths, in one or two no deaths are recorded at all. This means that the diabetes rates are subject to a considerable error,—addition or subtraction of a few cases would make all the difference to the calculated rate. Then again, although a small country, it is probable that Switzerland is by no means homogeneous, either racially or economically. The census of languages affords some, but by no means a complete, idea of this. Were the different racial elements uniformly scattered over the country, the heterogeneity would be of no great importance, but we can hardly assume this without an inquiry we have no means of undertaking. It will be seen then that the Swiss data are certainly not ideal figures for the study of this problem, although we do not question their utility and importance.

Italy.

In this case we escape from the difficulties attendant upon the use of a small population. The numbers are so large that even when subdivided into as many as 69 districts, we have in each case an absolute population sufficient to avoid any serious error with respect to the value of the calculated rates—so far as such error depends on the effects of random sampling. The questions of heterogeneity and the medical value of the records still remain. With regard to the former it is perhaps probable that the population is economically more uniform than is the case in industrialized communities such as England; with respect to racial elements we have no special information which enables us to offer an opinion.

So far as the accuracy of the records is concerned, we have little to go upon; in the mind of a statistician, the convenient and logical way in which the data are complied and published creates a prejudice in their favour, but this may only be a prejudice. It should, however, be remarked that the medical authorities in this country presumably recognise that the standard demanded by the Italian Government on admission to medical practice is not greatly different from our own, since Italy is the only European country with which complete reciprocity of practice exists. But this may be little evidence as to the real condition of medical science in Italy. On the facts before us we should be disposed to regard the Italian data as on the same footing as the English.

England.

We have used the following English data:

(1) The returns for registration counties as published by the Registrar General.

(2) The records for 1911 of towns having more than fifty thousand inhabitants, specially supplied to us with full material for calculating age and sex corrections by the courtesy of Dr Stevenson.

The English counties in respect of population are quite as satisfactory as Italy and superior to Switzerland. On the other hand the economic heterogeneity of the counties taken as a whole is enormous; indeed such as to preclude any expectation of satisfactory results when they are so taken. This may be to some extent avoided by using the Registrar General's classification into urban and rural counties, but we are doubtful how far this process is really successful. When we have formed more truly homogeneous groups the numbers are very small and any attempt to enlarge the number of districts upon which the coefficients are based introduces again the heterogeneity noted.

An additional source of difficulty is the enormous range of the populations; the range in the case of the other countries and cities is indeed considerable, but far less so than in the registration counties of England. The cities do not present the same difficulties, but here we come upon the other horn of the dilemma, since the absolute populations are too small to allow us to attach very much importance to the diabetes rates which are calculated from the returns of a single year.

With respect to the scientific reliability of the certified causes of death, it is unnecessary for us to offer any observations, the question being one that each reader can answer for himself. We may note that the unreliability of the rates in the case of the English cities is somewhat mitigated by the fact that we have a large number of separate cities—118 being available for tabulation.

This general account of the data will enable the reader to compare the substratum of our work with that upon which Maynard's conclusions were based. The general conclusions which we are disposed to draw may be stated as follows:

It is doubtful whether the presumed superiority of the Swiss statistics in regard to material and scientific accuracy of the returns really compensates for the paucity of numbers. From the latter point of view both the Italian and English data will sustain a comparison with Maynard's material. It has not, however, been possible to institute a completely valid comparison between European and American urban districts since, although our English cities are superior to Maynard's cities in point of numbers, they are inferior in size and in the validity of the rates—the latter being based upon the returns of a single year instead of on five years, as in the case of America. At the same time, were Maynard's correlation to mark an innate interrelationship of diabetes and cancer, a biological phenomenon *quod semper quod ubique quod ab omnibus* can be recognised, we should evidently expect to obtain a substantial agreement between the different series of results.

It will now be necessary to devote a few words to the question of statistical methods. As we stated above, Maynard employed either ${}_pr_{cd}$ or ${}_pr_{c'd'}$ where p = population, c and d the crude deaths and c' and d' the corrected deaths. Since the value of ${}_pr_{c'd'}$ might conceivably be affected by the introduction of spurious correlation, it can further be corrected by Pearson's method. With regard to the latter correction we may say (1) that in the case in which Pearson applied it to Maynard's coefficients no change was produced; (2) it is doubtful, in our opinion, whether the correction is ever likely to produce a substantial change. We ourselves endeavoured to form some idea of the magnitude of the spurious element introduced by correcting for age distribution by means of an empirical test.

Drawings were made from bags containing different proportions of red and white counters and the "rates," *i.e.* the proportions of red counters, in different drawings were corrected to a "standard population." A series of coefficients was calculated and the average result led us to believe that the spurious element introduced was unlikely to be a serious source of error in practical work. It would, however, be unscientific to attach too much importance either to general ideas or to the result of a single test; we have in some cases applied Pearson's correction and have never succeeded in altering the coefficients to an extent which would affect the conclusions based upon them.

A more important matter is connected with the difference between ${}_pr_{cd}$ and ${}_pr_{\frac{c}{p}\frac{d}{p}}$. In theory, under certain specified conditions, these coefficients should be identical and it can be proved that they *are* identical when the standard deviations of the various characters are small in comparison with their respective means. This condition is not fulfilled by any of our series and there is consequently a marked divergence between the two coefficients in some cases. The only instance in which the difference is such that it might affect the

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

reasoning based upon the correlations is that of English registration counties.

We have elsewhere recorded our grounds for believing that in cases of divergences the coefficient ${}_{p}r_{\frac{c}{p}\frac{d}{p}}$ is to be preferred[1], but, since the matter may be regarded as controversial, we give both coefficients in those cases which reveal important differences.

Some other peculiarities of the present inquiry are (1) in view of the close similarity in age distribution of both diseases it is plain that, if the crude rates are not sensibly correlated, it is improbable that the corrected rates will be. Consequently, if the correlation of crude rates is not significant, it is hardly worth while correlating corrected rates; (2) if it is desired to work with corrected rates, it is sufficient to calculate the age correction factor for one of the two diseases, since the two factors are found to be very closely correlated; (3) we have in most cases operated upon males alone, but experience suggests that it is not necessary to correct for sex distribution. As an example of this we have calculated the partial correlation between deaths from diabetes and deaths from cancer for population constant using 118 English towns, in the case of all persons and in that of males only. The corresponding figures are ·3892 and ·3820 (with respect to the values of these correlations it should be remarked that they are based upon crude figures uncorrected for age distribution).

ANALYSIS OF RESULTS.

Switzerland (Tables V and VI).

Taking the 25 Cantons as our subdivisions and confining ourselves to males, we first correlated the absolute numbers of deaths from cancer and diabetes respectively for constant population, and found $-·1741 \pm ·1310$. If in addition to population we keep the absolute numbers of deaths of lunatics and deaths from cardiac disease constant, we reach $-·1531 \pm ·1317$. Neither value is definitely significant. Using Maynard's method of correlating corrected numbers of deaths, we have $-·0337 \pm ·1347$. Simply correlating the corrected rates with population constant we have $-·1533 \pm ·1182$. If the rates are corrected on the bases of the deaths at ages in the subpopulations (the previous corrections were based on crude total death-rates in subpopulations and the age distributions of the same without reference to deaths at ages in the subpopulations)

[1] *Journ. Roy. Stat. Soc.* Febr. 1914.

the correlation is $-\cdot1666 \pm \cdot1312$ and, keeping cardiac diseases and deaths of lunatics constant, we have $-\cdot0394 \pm \cdot1347$. None of these values are significant. The same material was then grouped into 23 districts by putting the two Appenzell Cantons together and the two Unterwalden Cantons together. In this case the corrected rates were correlated—the populations being kept constant, the result was $-\cdot1416 \pm \cdot1322$, also insignificant. Twenty Cantons were then taken, by the

TABLE V. SWITZERLAND[1].

Correlations between the mortality from cancer and diabetes for Switzerland, based upon an average of 5 years (1901–1905).

Variables	Correlations
Crude absolute numbers of deaths. Males. 25 Cantons.	
Cancer and diabetes	+·7612 ± ·0567
Cancer and population	+·9524 ± ·0125
Diabetes and population	+·8303 ± ·0419
Cancer and diabetes with population constant ...	−·1741 ± ·1308
Cancer and cardiac disease	+·9675 ± ·0086
Cancer and lunacy	+·8162 ± ·0450
Diabetes and cardiac disease	+·8170 ± ·0449
Diabetes and lunacy	+·8662 ± ·0337
Cardiac disease and lunacy	+·8629 ± ·0345
Cancer and diabetes with cardiac disease and lunacy constant	−·1531 ± ·1317
Corrected absolute numbers of deaths[2]. *Males. 25 Cantons.*	
Cancer and diabetes	+·7886 ± ·0510
Cancer and population...	+·9696 ± ·0081
Diabetes and population	+·8183 ± ·0446
Cancer and diabetes with population constant ...	−·0337 ± ·1347

[1] The number of deaths from cancer and the populations of the different Cantons were obtained from "Statistique de la Suisse," *Mouvement de la Population de la Suisse*, while the number of deaths from diabetes as well as the death-rate at ages in the sub-populations for both cancer and diabetes were obtained direct from the Bureau Fédéral de Statistique.

[2] Owing to the different age distribution in the various Cantons, the crude number of deaths or the crude death-rates have to be multiplied by appropriate correction factors. These correction factors can be calculated by two methods, (1) when the death-rates at ages in all the subpopulations are known, and (2) when the death-rates at ages in the subpopulations are not known. For the Swiss data only could the death-rates at ages in the subpopulations be obtained and in certain cases (marked *) the correction factors have been calculated by the first method. In all other cases the second method was used. As a matter of fact the final values obtained by the two methods do not differ markedly. For a full description of the two methods see *An Introduction to the Study of Statistics*, by G. Udny Yule, pp. 223–225.

Corrected death-rates. Males. 25 Cantons.

Cancer and diabetes	− ·1756 ± ·1307
Cancer and population...	− ·3011 ± ·1227
Diabetes and population	+ ·1001 ± ·1335
Cancer and diabetes with population constant ...	− ·1533 ± ·1182

Corrected death-rates. Males. 25 Cantons.*

Cancer and diabetes	− ·1666 ± ·1312
Cancer and cardiac disease	+ ·2039 ± ·1279
Cancer and lunacy	− ·4624 ± ·1061
Diabetes and cardiac disease	+ ·4538 ± ·1071
Diabetes and lunacy	+ ·5441 ± ·0950
Cardiac disease and lunacy	+ ·3765 ± ·1158
Cancer and diabetes with cardiac disease and lunacy constant	− ·0394 ± ·1347

Corrected death-rates. Males. 23 Districts,* the two Appenzells and the two Unterwalds being taken together.

Cancer and diabetes	− ·1529 ± ·1374
Cancer and population	+ ·2015 ± ·1349
Diabetes and population	+ ·0722 ± ·1399
Cancer and diabetes with population constant ...	− ·1416 ± ·1322

Corrected death-rates. Males. 20 Cantons,* 5 Cantons containing towns of more than 45,000 inhabitants being omitted.

Cancer and diabetes	− ·1501 ± ·1474
Cancer and cardiac disease	+ ·3145 ± ·1359
Cancer and lunacy	− ·4871 ± ·1150
Diabetes and cardiac disease	+ ·2122 ± ·1440
Diabetes and lunacy	+ ·3989 ± ·1268
Cardiac disease and lunacy	+ ·3349 ± ·1339
Cancer and diabetes with cardiac disease and lunacy constant	+ ·0227 ± ·1507

Crude death-rates. Males. 18 Towns with more than 10,000 inhabitants.

Cancer and diabetes	− ·2407 ± ·1498

Crude absolute numbers of deaths. Females. 25 Cantons.

Cancer and diabetes	+ ·7517 ± ·0587
Cancer and population	+ ·9603 ± ·0105
Diabetes and population	+ ·7235 ± ·0643
Cancer and diabetes with population constant ...	+ ·2957 ± ·1231

Corrected death-rates. Males and Females. 25 Cantons.*

Cancer and diabetes	− ·0722 ± ·1342

omission of Cantons with towns of more than 45,000 inhabitants. Here also the correlations were inappreciable.

Lastly we correlated the rates corrected for ages (using deaths at ages in subpopulations) but without separating the sexes, and found − ·0722 ± ·1342.

TABLE VI. SWITZERLAND.

Means and standard deviations for the Swiss data based upon an average of 5 years (1901–1905).

Variable	Mean	Standard deviation
Crude absolute numbers of deaths. Males. 25 Cantons.		
Cancer	82·656	77·865
Diabetes... ...	4·776	5·003
Cardiac disease ...	128·672	133·440
Lunacy	5·800	6·072
Population (1900) ...	65,081·0	66,100·0
Corrected absolute numbers of deaths. Males. 25 Cantons.		
Cancer	83·161	82·129
Diabetes... ...	4·893	5·293
Corrected death-rates[1]. *Males. 25 Cantons.*		
Cancer	1418·120	458·729
Diabetes... ...	69·496	55·885
Corrected death-rates. Males. 25 Cantons.*		
Cancer	1425·306	464·273
Diabetes... ...	70·135	57·581
Cardiac disease ...	1868·647	390·586
Lunacy	79·163	41·692
Corrected death-rates. Males. 23 Districts, the two Appenzells and the two Unterwalds being taken together.*		
Cancer	1357·581	348·180
Diabetes... ...	72·565	56·349
Population (1900) ...	70,740·2	54,345·7
Corrected death-rates. Males. 20 Cantons, 5 Cantons containing towns with more than 45,000 inhabitants being omitted.*		
Cancer	1473·352	502·165
Diabetes... ...	53·779	38·358
Cardiac disease ...	1785·658	365·184
Lunacy	71·376	39·629
Crude death-rates. Males. 13 towns with more than 10,000 inhabitants.*		
Cancer	1267·78	287·36
Diabetes... ...	100·00	38·884
Crude absolute numbers of deaths. Females. 25 Cantons.		
Cancer	84·224	82·607
Diabetes... ...	3·376	3·530
Population (1900) ...	67,536·7	67,147·0
Corrected Death-rates. Males and Females. 25 Cantons.*		
Cancer	1261·505	354·571
Diabetes	55·850	41·507

[1] Per 1,000,000 living. * See foot-note (2) to Table V.

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

We have also worked out the correlation between the crude *rates* of cancer and diabetes in the 19 towns of Switzerland which contain more than 10,000 inhabitants. The result is $-\cdot2407 \pm \cdot1498$.

From a consideration of these results which are given with further details in the tables, it will be seen that no grouping of the material or variation in the method of calculation, *i.e.* the use of rates instead of absolute numbers, will produce a correlation coefficient which is definitely significant with regard to its probable error. Hence we must conclude that so far as Switzerland is concerned no correlation between the two rates can be shown to exist.

Italy (Tables VII and VIII).

We have analysed the Italian data in several different ways, as will be seen from the tables. As has been observed before, this material was in point of numbers probably the most satisfactory at our command. It will be seen that in the case of the 69 smaller administrative units (not one of these subdivisions contained less than 100,000 inhabitants), whether we use the method of rates or of the partials based on absolute numbers both in the case of crude or in that of age corrected figures, the result is substantially the same—there is no distinct evidence of a significant association. The utmost that can be contended is that a small positive correlation exists. When we deal with the larger units, the method of calculation is more influential on the result and if we attach more importance to coefficients based upon rates we should argue that some positive correlation exists. Even here, however, in the most striking instance, the coefficient is not quite thrice its probable error. The diminution in intensity of the correlation as we pass from larger to smaller units is what might have been predicted from Maynard's results (compare his coefficients for states with those yielded by cities), but the absolute values are strikingly different in the two cases. The 69 subdistricts are, it is to be presumed, more heterogeneous than Maynard's cities, since in our case we have mixed urban with rural areas. But it is difficult to suppose that the differences between the coefficients in the two cases can be explained entirely in this way, because the 16 provinces can hardly be more heterogeneous than the American states and are most likely less so. But, even here we find a very marked difference. We have already referred to the question of material accuracy and have admitted our inability to appraise the data from that point of view. But it does not seem reasonable to suppose

TABLE VII. ITALY[1].

Correlation between mortality from cancer and from diabetes for Italy. Males and Females.

Variables	Correlation
A. 16 DIVISIONS.	
(1) *Crude absolute number of deaths in* 1905.	
Deaths from cancer and from diabetes	+ ·7221 ± ·0807
Deaths from cancer and population	+ ·8767 ± ·0390
Deaths from diabetes and population	+ ·8254 ± ·0537
Deaths from cancer and from diabetes with population constant	− ·0055 ± ·1686
(2) *Crude absolute number of deaths in* 1906.	
Deaths from cancer and from diabetes	+ ·8094 ± ·0582
Deaths from cancer and population	+ ·8796 ± ·0382
Deaths from diabetes and population	+ ·9276 ± ·0235
Deaths from cancer and from diabetes with population constant	− ·0365 ± ·1684
(3) *Crude death-rates based upon an average of* 5 *years* (1905–1909).	
Death-rate from cancer and from diabetes ...	+ ·3356 ± ·1496
Death-rate from cancer and population	+ ·3685 ± ·1457
Death-rate from diabetes and population	− ·0728 ± ·1677
Death-rate from cancer and from diabetes with population constant	+ ·3909 ± ·1429
(4) *Corrected absolute number of deaths based upon an average of* 5 *years* (1905–1909).	
Deaths from cancer and from diabetes	+ ·8328 ± ·0517
Deaths from cancer and population	+ ·8667 ± ·0420
Deaths from diabetes and population	+ ·9250 ± ·0243
Deaths from cancer and from diabetes with population constant	+ ·1640 ± ·1641
(5) *Corrected death-rates based upon an average of* 5 *years* (1905–1909).	
Death-rate from cancer and from diabetes... ...	+ ·3875 ± ·1602
Death-rate from cancer and population	+ ·2929 ± ·1542
Death-rate from diabetes and population	− ·1125 ± ·1665
Death-rate from cancer and diabetes with population constant	+ ·4425 ± ·1356
B. 69 COMPARTMENTS.	
(1) *Crude absolute number of deaths based upon an average of* 2 *years* (1905 & 1906).	
Deaths from cancer and from diabetes	+ ·7845 ± ·0312
Deaths from cancer and population	+ ·8449 ± ·0232
Deaths from diabetes and population	+ ·8857 ± ·0175
Deaths from cancer and from diabetes with population constant	+ ·1458 ± ·0795
(2) *Corrected absolute number of deaths based upon an average of* 2 *years* (1905 & 1906).	
Deaths from cancer and from diabetes	+ ·7951 ± ·0299
Deaths from cancer and population	+ ·8542 ± ·0220
Deaths from diabetes and population	+ ·8867 ± ·0174
Deaths from cancer and from diabetes with population constant	+ ·1566 ± ·0792
(3) *Corrected death-rates based upon an average of* 2 *years* (1905 & 1906).	
Death-rate from cancer and from diabetes... ...	+ ·1900 ± ·0783
Death-rate from cancer and population	− ·0536 ± ·0810
Death-rate from diabetes and population	+ ·2904 ± ·0744
Death-rate from cancer and from diabetes with population constant	+ ·2151 ± ·0774

[1] Data obtained from "Statistica delle Cause di Morte," *Direzione Generale Della Statistica*, for the years in question.

that the Italian standard is markedly inferior to the American, it may even be higher. It is, in our opinion, very difficult to believe that our failure to obtain the same or approximately the same results as Maynard's can be accounted for by shortcomings of the data. The question of statistical method, on the other hand, does not arise, as the findings of

TABLE VIII. ITALY.

Means and standard deviations for the Italian data.
Males and Females.

Variable	Mean	Standard deviation
A. 16 DIVISIONS.		
(1) *Crude absolute number of deaths in* 1905.		
Deaths from cancer	1209·250	912·763
Deaths from diabetes	82·625	47·809
Population (calculated to the middle of 1905)	2,085,135	1,182,470
(2) *Crude absolute number of deaths in* 1906.		
Deaths from cancer	1290·810	968·413
Deaths from diabetes	87·125	51·327
Population (calculated to the middle of 1905)	2,085,135	1,182,470
(3) *Crude death-rates per* 1,000,000 *living. Average of* 5 *years* (1905–1909).		
Death-rate from cancer	576·862	197·180
Death-rate from diabetes	43·809	13·355
Population (calculated to the middle of 1907)	2,111,005	1,195,712
(4) *Corrected absolute number of deaths. Average of* 5 *years* (1905–1909).		
Deaths from cancer	1358·593	995·924
Deaths from diabetes	94·589	51·786
Population (calculated to the middle of 1907)	2,111,005	1,195,712
(5) *Corrected death-rate per* 1,000,000 *living. Average of* 5 *years* (1905–1909).		
Death-rate from cancer	607·688	215·753
Death-rate from diabetes	45·665	13·412
Population (calculated to the middle of 1907)	2,111,005	1,195,712
B. 69 COMPARTMENTS.		
(1) *Crude absolute number of deaths. Average of* 2 *years* (1905 *&* 1906).		
Deaths from cancer	289·862	213·897
Deaths from diabetes	19·681	16·434
Population (calculated to the middle of 1905)	483,510	250,226
(2) *Corrected absolute number of deaths. Average of* 2 *years* (1905 *&* 1906).		
Deaths from cancer	290·108	223·008
Deaths from diabetes	19·912	17·122
Population (calculated to the middle of 1905)	483,320	281,242
(3) *Corrected death-rate per* 1,000,000 *living. Average of* 2 *years* (1905 *&* 1906).		
Death-rate from cancer	607·188	222·822
Death-rate from diabetes	38·791	14·252
Population (calculated to the middle of 1905)	483,320	281,242

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

identical methods are furnished for comparison. The conclusion seems to be almost inevitable that the cause or causes which produce a high correlation between the rates in America, either do not operate or are overmastered by some other factors in Italy. It would seem that this part of our investigation strengthens that dealing with Swiss data.

England (Tables IX, X, XI, XII, XIII and XIV).

Our English data fall into two classes, viz. the large towns and the registration counties. In the case of the larger towns we were able by excluding a few very large cities which exceeded all the others in point of population by a considerable margin, to obtain a group of 118 urban communities, which contained a sufficient number of observations to allow one to form correlation tables. These tables were treated, as will be seen in the schedule of results (Table IX), by numerous methods and

TABLE IX. ENGLISH TOWNS.

Correlation between mortality from cancer and from diabetes in 118 English towns with populations between 50,000 and 375,000[1]. *Males and Females.*

Variables	r	η	$\frac{\sqrt{N}}{\cdot 67449} \times \frac{1}{2}\sqrt{\eta^2 - r^2}$
Crude absolute numbers of deaths using Pearson's method of correcting for age distribution[2].			
Deaths from cancer and from diabetes ...	+ ·8925	·9174	1·710
	± ·0126	·9094	1·405
Deaths from cancer and population	+ ·9423	·9511	1·044
	± ·0070	·9526	1·273
Deaths from diabetes and population ...	+ ·8820	·9003	1·454
	± ·0138	·9080	1·736
Deaths from cancer and from diabetes with population constant	+ ·3892	—	—
	± ·0527		
Deaths from cancer and cancer corrective factor...	− ·1426	—	—
	± ·0608		

[1] The data, which are given in full in Table XI, were kindly supplied to us by Dr Stevenson.

[2] This method consists in calculating the partial correlation between the mortality from cancer and from diabetes with the cancer and diabetes corrective factors constant. As the correlation between the two corrective factors is always very high (·9984 and ·9825 were two of the values found during the present investigation) the cancer corrective factors can be used for the diabetes data, thus effecting a considerable reduction in the number of correlation coefficients to be calculated. See "On the Correlation of Death-Rates" by K. Pearson, F.R.S., assisted by Alice Lee, D.Sc., and Ethel M. Elderton, *Journal of the Royal Statistical Society*, Vol. LXXIII. p. 534.

Deaths from diabetes and cancer corrective factor	− ·1449	—	—
	± ·0608		
Population and cancer corrective factor ...	+ ·0343	—	—
	± ·0620		
Deaths from cancer and from diabetes with population and cancer corrective factor constant	+ ·2461	—	—
	± ·0583		
Crude death-rates using Pearson's method of correcting for age distribution.			
Death-rate from cancer and from diabetes ...	+ ·3564	·4989	2·731
	± ·0542	·3892	1·259
Death-rate from cancer and population ...	+ ·0430	·2364	1·872
	± ·0620	·3681	2·944
Death-rate from diabetes and population ...	+ ·0031	·2166	1·744
	± ·0621	·1435	1·156
Death-rate from cancer and from diabetes with population constant	+ ·3566	—	—
	± ·0542		
Death-rate from cancer and cancer corrective factor	− ·7485	—	—
	± ·0273		
Death-rate from diabetes and cancer corrective factor	− ·4441	—	—
	± ·0498		
Death-rate from cancer and from diabetes with population and cancer corrective factor constant	+ ·0385	—	—
	± ·0620		
Corrected absolute numbers of deaths.			
Deaths from cancer and from diabetes ...	+ ·8799	·9048	1·697
	± ·0140	·9163	2·059
Deaths from cancer and population	+ ·9728	·9795	·924
	± ·0033	·9782	·829
Deaths from diabetes and population ...	+ ·8909	·9106	1·520
	± ·0128	·9279	2·090
Deaths from cancer and from diabetes with population constant	+ ·1259	—	—
	± ·0611		
Corrected death-rates.			
Death-rate from cancer and from diabetes ...	+ ·0475	·3569	2·849
	± ·0619	·2644	2·095
Death-rate from cancer and population ...	+ ·1218	·2694	1·935
	± ·0529	·4466	3·460
Death-rate from diabetes and population ...	+ ·0462	·2106	1·655
	± ·0620	·4264	3·413
Death-rate from cancer and from diabetes with population constant	+ ·0438	—	—
	± ·0620		
Corrected absolute numbers of deaths. Males only.			
Deaths from cancer and from diabetes ...	+ ·7978	—	—
	± ·0226		
Deaths from cancer and population	+ ·9167	—	—
	± ·0099		
Deaths from diabetes and population ...	+ ·7625	—	—
	± ·0260		
Deaths from cancer and from diabetes with population constant	+ ·3820	—	—
	± ·0530		

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

allowance was made by the use of Pearson's corrective process, for the possibility of spurious correlation. The differences between the values of coefficients obtained in different ways, if not entirely negligible (that the various regressions are non-linear was proved by a special investigation which is detailed in another paper), are not sufficient to lead to difficulties of interpretation. The general result is to suggest that some degree of correlation between the rates exists, but that it is of an order wholly different from that found by Maynard. Our material, however, differs from his in two ways, one favourable to its value, the other

TABLE X. English Towns.

Means and standard deviations for the English Towns, 1911. *Males and Females.*

Variable	Mean	Standard deviation
Crude number of deaths from cancer	119·915	78·910
Crude number of deaths from diabetes	11·987	8·320
Crude death-rate[1] from cancer	986·017	214·948
Crude death-rate from diabetes	98·453	37·258
Corrected number of deaths from cancer	116·780	74·103
Corrected number of deaths from diabetes	11·644	8·089
Corrected death-rate from cancer	995·339	127·099
Corrected death-rate from diabetes	95·254	32·819
Cancer corrective factor[2]	·98750	·14008
Population (1911)	122,458	72,112
Corrected number of deaths from cancer. Males only...	52·701	34·636
Corrected number of deaths from diabetes. Males only	5·841	4·143
Population. Males only (1911)...	58,501	34,751

[1] Per 1,000,000 living.

[2] Age distribution given in the 1911 Census was used for the calculation of the correction factors.

decidedly unfavourable. In our favour is the fact that the number of separate observations is nearly three times as great as that of Maynard. Against us is the necessity which compelled us to base the calculations on the records of deaths in a single year. The obvious objection to this is the uncertainty attaching to the individual records. The probable error is based upon the number of separate observations, *i.e.* the number of cities, but takes no account of the question as to the value of the separate records.

So far as the probable error calculation is concerned, 100 cities of 5000 inhabitants each would give, by the method we are using, the same probable error as 100 cities of 50,000 inhabitants apiece if the

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

TABLE XI. *Deaths from cancer and from diabetes during* 1911 *in* 126[1] *English towns with populations of more than* 50,000.

Town	Deaths from diabetes		Deaths from cancer	
	Males	Females	Males	Females
Barrow-in-Furness	0	4	15	25
Bath	2	2	21	49
Birkenhead	5	5	59	64
Birmingham[2]	26	23	206	314
Blackburn	5	8	40	77
Blackpool	6	5	34	47
Bolton	7	11	69	86
Bootle	5	3	32	42
Bournemouth	4	5	39	61
Bradford	19	20	141	212
Brighton	6	10	76	113
Bristol[2]	19	24	149	233
Burnley	4	4	35	50
Bury	2	3	16	30
Cardiff	7	5	74	102
Chester	1	3	16	24
Coventry	6	3	30	42
Croydon	6	7	78	109
Derby	6	9	52	77
Devonport	2	5	33	37
Dudley	4	2	17	17
Eastbourne	5	3	20	38
Gateshead	7	3	47	42
Gloucester	2	2	21	19
Great Yarmouth	4	1	18	42
Grimsby	5	4	28	32
Halifax	8	6	60	58
Hastings	4	4	47	63
Huddersfield	4	5	49	80
Ipswich	6	4	35	51
Kingston-on-Hull	14	14	117	171
Leeds[2]	31	22	191	251
Leicester	11	17	81	154
Lincoln	4	1	36	35
Liverpool[2]	23	31	311	392
Manchester[2]	50	27	342	430
Merthyr Tydfil	3	1	32	27
Middlesborough	4	5	36	43
Newcastle[2]	18	13	116	145
Newport (Mon.)	1	4	23	47
Northampton	6	6	35	54
Norwich	6	7	64	64
Nottingham	15	11	104	172
Oldham	7	4	62	86
Oxford	2	3	30	43
Plymouth	7	5	60	63
Portsmouth	12	11	88	108
Preston	83	7	47	64
Reading	9	4	32	48
Rochdale	7	6	52	63
Rotherham	5	3	25	25
St Helens	7	4	28	32
Salford	17	15	105	117
Sheffield[2]	19	15	181	203
Smethwick	5	2	18	39
Southampton	6	8	62	66
Southport	5	5	27	40
S. Shields	4	11	44	45
Stockport	3	13	51	68
Stoke-on-Trent	9	9	79	96
Sunderland	5	10	53	70
Swansea	3	1	42	58
Tynemouth	0	6	23	27
Walsall	1	0	26	41
Warrington	0	1	26	26
W. Bromwich	2	4	31	36
W. Ham	9	15	128	138
W. Hartlepool	5	3	14	37
Wigan	6	5	14	34
Wolverhampton	4	5	48	50
Worcester	4	3	20	40
York	8	5	27	47
Wallasey	1	6	31	59
Darlington	1	3	27	30
Stockton-on-Tees	0	2	17	22
East Ham	4	6	53	56
Ilford	1	1	25	35
Leyton	2	7	50	60
Southend-on-Sea	3	9	27	43
Walthamstow	1	1	49	57
Gillingham	1	2	25	21
Acton	5	0	19	33
Ealing	3	6	31	45
Edmonton	1	0	30	29
Enfield	2	4	23	26
Hornsey	3	4	29	71
Tottenham	5	7	47	65
Willesden	4	8	39	77
Handsworth	2	4	28	29
Wimbledon	5	0	18	30
Aston Manor	3	2	26	40
Swindon	1	3	17	22
King's Norton	5	2	19	36
Barnsley	3	2	23	28
Dewsbury	4	7	23	31
Wakefield	3	0	25	46
Aberdare	2	2	11	24
Rhondda	6	4	37	39
Battersea	11	9	76	91
Bermondsey	8	6	63	56
Bethnal Green	8	3	57	64
Camberwell	14	10	124	145
Chelsea	3	2	37	55
Deptford	2	6	52	75
Finsbury	2	4	44	38
Fulham	9	7	80	94
Greenwich	3	1	41	40
Hackney	12	14	94	133
Hammersmith	9	9	66	64
Hampstead	8	5	26	65
Holborn	5	2	27	30
Islington	14	15	152	184
Kensington	7	14	91	116
Lambeth[2]	12	16	162	175
Lewisham	13	14	77	104
Paddington	7	8	71	96
Poplar	9	6	86	65
St Marylebone	5	3	90	77
St Pancras[2]	7	15	122	153
Shoreditch	5	2	51	40
Southwark	5	9	98	113
Stepney	10	19	137	108
Stoke Newington	3	5	30	34
Wandsworth	11	14	121	204
Westminster	12	5	107	104
Woolwich	12	6	65	64

[1] In calculating the correlations given in Table IX only 118 towns and boroughs were used; 8 towns with populations of over 375,000 were omitted.

[2] These towns and boroughs with population over 200,000 and having medical schools were used for the correlation given in Table I (2).

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

TABLE XII. ENGLAND AND WALES.

Mean annual death-rates per million living, 1901–1910.

	Cancer		Diabetes	
Age	Males	Females	Males	Females
0–	36	29	4	5
5–	18	13	10	10
10–	17	15	19	20
15–	31	27	36	27
20–	53	39	46	35
25–	109	170	59	51
35–	414	846	79	63
45–	1549	2321	160	129
55–	3904	4410	415	357
65–	6683	6658	731	574
75 and upwards	7874	7901	720	473
All ages	773	1027	103	90

TABLE XIII. ENGLISH COUNTIES*.

Correlation between mortality from cancer and from diabetes for English registration counties based upon an average of 6 *years* (1905–1910). *Males.*

Variables	Correlation
A. 41 *English Counties; Rural, "Mixed," and Urban.*	
(1) *Crude absolute number of deaths; using Pearson's method of correcting for age distribution*[1].	
Deaths from cancer and from diabetes	+ ·9741 ± ·0054
Deaths from cancer and population	+ ·9852 ± ·0031
Deaths from diabetes and population	+ ·9893 ± ·0022
Deaths from cancer and from diabetes with population constant	− ·0265 ± ·1053
Deaths from cancer and cancer corrective factor ...	+ ·5423 ± ·0744
Deaths from diabetes and cancer corrective factor ...	+ ·5617 ± ·0721
Population and cancer corrective factor	+ ·6153 ± ·0655
Deaths from cancer and from diabetes with population and cancer corrective factor constant	− ·2752 ± ·0974
(2) *Crude death-rates; using Pearson's method of correcting for age distribution.*	
Deaths from cancer and from diabetes	+ ·6635 ± ·0590
Deaths from cancer and population	− ·2655 ± ·0979
Deaths from diabetes and population	− ·3197 ± ·0946
Deaths from cancer and from diabetes with population constant	+ ·6334 ± ·0631
Deaths from cancer and cancer corrective factor ...	− ·6445 ± ·0616
Deaths from diabetes and cancer corrective factor ...	− ·6401 ± ·0622
Deaths from cancer and from diabetes with population and cancer corrective factor constant	+ ·4135 ± ·0873

* Data obtained from the *Annual Reports of the Registrar General of England and Wales* (1905–1910).

[1] See foot-note (2) to Table IX.

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

B. 32 *English Counties; Rural and "Mixed."*

(1) *Crude absolute number of deaths; using Pearson's method of correcting for age distribution.*

Deaths from cancer and from diabetes	+ ·9662 ± ·0079
Deaths from cancer and population	+ ·9517 ± ·0112
Deaths from diabetes and population	+ ·8973 ± ·0232
Deaths from cancer and from diabetes with population constant	+ ·8286 ± ·0374
Deaths from cancer and cancer corrective factor ...	+ ·6125 ± ·0745
Deaths from diabetes and cancer corrective factor ...	+ ·5496 ± ·0832
Population and cancer corrective factor	+ ·7558 ± ·0511
Deaths from cancer and from diabetes with population and cancer corrective factor constant	+ ·7806 ± ·0466
Deaths from cancer and deaths from "other causes"[1] ...	+ ·9026 ± ·0221
Deaths from diabetes and deaths from "other causes"[1]	+ ·8546 ± ·0322
Population and deaths from "other causes"	+ ·9826 ± ·0041
Cancer corrective factor and deaths from "other causes"[2]	+ ·7680 ± ·0489
Deaths from cancer and from diabetes keeping population, cancer corrective factor and deaths from "other causes" constant	+ ·7871 ± ·0454

(2) *Crude death-rates; using Pearson's method of correcting for age distribution.*

Death-rate from cancer and from diabetes	+ ·6600 ± ·0673
Death-rate from cancer and population	− ·6151 ± ·0741
Death-rate from diabetes and population	− ·5617 ± ·0816
Death-rate from cancer and from diabetes with population constant	+ ·4821 ± ·0915
Death-rate from cancer and cancer corrective factor ...	− ·7786 ± ·0470
Death-rate from diabetes and cancer corrective factor ...	− ·7331 ± ·0552
Death-rate from cancer and from diabetes with population and cancer corrective factor constant	+ ·2068 ± ·1141

(3) *Corrected death-rates.*

Death-rate from cancer and from diabetes	+ ·2376 ± ·1125
Death-rate from cancer and population	− ·0516 ± ·1189
Death-rate from diabetes and population	− ·3293 ± ·1063
Death-rate from cancer and from diabetes with population constant	+ ·2339 ± ·1127

C. 9 *Urban Counties.*

(1) *Crude absolute number of deaths.*

Deaths from cancer and from diabetes[3]	+ ·9617 S.D. ·0266
Deaths from cancer and population	+ ·9693 S.D. ·0214
Deaths from diabetes and population	+ ·9965 S.D. ·0025
Deaths from cancer and from diabetes with population constant	− ·2020 S.D. ·3391

[1] *I.e.* deaths from all causes except cancer and diabetes.

[2] Correlation between "other deaths" correction factors and cancer correction factors is very high (+ ·9938), so that again only the cancer correction factors need be used.

[3] As the number of observations is small (9), the usual values of the S.D. and P.E. of r are not applicable. The values given of the S.D. of r have been calculated from the formula $\sigma_r = \frac{1-r^2}{\sqrt{n-1}}$; see "On the Probable Error of the Correlation Coefficient," by H. E. Soper, M.A. (*Biom.* IX. p. 91).

(2) *Corrected absolute number of deaths.*

Deaths from cancer and from diabetes	+·9638	S.D. ·0251
Deaths from cancer and population	+·9761	S.D. ·0167
Deaths from diabetes and population	+·9914	S.D. ·0061
Deaths from cancer and from diabetes with population constant	−·1358	S.D. ·3470

D. 13 *Rural Counties.*

(1) *Crude absolute number of deaths.*

Deaths from cancer and from diabetes	+·9866	S.D. ·0077
Deaths from cancer and population	+·9911	S.D. ·0051
Deaths from diabetes and population	+·9817	S.D. ·0105
Deaths from cancer and from diabetes with population constant	+·5375	S.D. ·2058

(2) *Corrected absolute number of deaths.*

Deaths from cancer and from diabetes	+·9868	S.D. ·0076
Deaths from cancer and population	+·9949	S.D. ·0029
Deaths from diabetes and population	+·9847	S.D. ·0088
Deaths from cancer and from diabetes with population constant	+·4039	S.D. ·2416

coefficients of correlation were the same in the two cases. But no one would seriously argue that the two results were really of equal value since with such a disease as diabetes a few cases wrongly diagnosed or improperly included in the local records might greatly affect the rate. To what extent these sources of error deprive our results of value we cannot say. It is obviously impossible to argue that the correlations are probably as high as in the American cities and would have so appeared if we had had a wider range of time upon which to base an average, but it is impossible with the data at our command to *prove* that this circumstance has not affected our values.

If we now turn to the results obtained from registration counties, we are faced with an almost insuperable difficulty of drawing trustworthy conclusions. Here the method of calculation makes an enormous difference to the results.

We will begin with the analysis of 32 rural and semi-rural counties. The object of considering this group separately is that from the point of view of industrial character this should be the most homogeneous group of reasonable size we can form from registration counties. This is not to say, of course, that it is homogeneous—far from it, but we can do no better without reducing the number of available districts to a value too small to render the calculations of any importance. Let us first take the results obtained when the methods adopted by Maynard and Pearson, viz. operating upon absolute numbers, are employed. We first correlated absolute deaths from the two diseases with population and cancer corrective factor constant (it was unnecessary to use more than one age correction factor since the diabetes and cancer factors were very

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

TABLE XIV. ENGLISH COUNTIES.

Means and standard deviations for the English counties data, based upon an average of 6 years (1905–1910). Males.

Variable	Mean	Standard deviation
A. 41 *English Counties, Urban, "Mixed," and Rural.*		
Crude deaths from cancer	313·967	406·606
Crude deaths from diabetes	41·984	50·711
Crude death-rate from cancer[1]	870·447	138·942
Crude death-rate from diabetes	118·983	21·811
Population[2]	380,844	475,036
Cancer corrective factor[3]	·93541	·13900
B. 32 *English Counties, "Mixed" and Rural.*		
Crude deaths from cancer	192·771	111·722
Crude deaths from diabetes	26·229	14·403
Crude deaths from "other causes"	3144·328	2257·073
Crude death-rate from cancer	898·906	132·703
Crude death-rate from diabetes	123·281	22·456
Corrected death-rate from cancer	795·625	72·188
Corrected death-rate from diabetes	112·969	15·325
Population	228,488	154,472
Cancer corrective factor	·89855	·11661
C. 9 *Urban Counties.*		
Crude deaths from cancer	744·889	686·197
Crude deaths from diabetes	98·000	83·410
Corrected deaths from cancer	833·278	813·713
Corrected deaths from diabetes	109·852	101·005
Population	922,556	753,111
D. 13 *Rural Counties.*		
Crude deaths from cancer	123·372	65·306
Crude deaths from diabetes	18·167	10·560
Corrected deaths from cancer	98·885	53·015
Corrected deaths from diabetes	14·577	8·535
Population	131,715	73,873

[1] Death-rate per 1,000,000 living.

[2] The populations are all calculated for the year 1907 according to the method described on pages xi and xii of the 73rd *Report of the Registrar General of England and Wales* using the appropriate populations in 1901 and 1911 as basis.

[3] The age distribution given in the 1901 Census was used for the calculation of the correction factors.

highly correlated; this was also true, curiously enough, when the age correction factor for deaths due to diseases other than diabetes and cancer was calculated; it was very highly correlated with the cancer correction factor). The result is a very substantial correlation. We then went a step further and introduced another variable, namely the

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

deaths from causes other than diabetes and cancer. This did not produce any change in the value of the partial correlation which is of precisely the same order as found by Maynard in the case of American cities (·7871 ± ·0454).

At first sight this would seem to prove that Maynard's finding is directly applicable to English registration counties and that the failure to obtain similar values in the case of other data should be disregarded, but a very curious peculiarity was revealed. We have above referred to the question of method and noted that in some cases the employment of a death-rate as variable instead of the absolute number of deaths makes a difference in the value of the partial correlation coefficient. We also remarked that in the case of Maynard's own material re-calculation of his results by the other method, viz. the employment of rates, did not in fact substantially modify the values of his coefficients. But in the case of these 32 counties the difference is very great. Thus if we use as variables the crude death-rate from cancer, the crude death-rate from diabetes, the population and the cancer age corrective factor, the partial correlation between the first two variables is only ·2068 ± ·1141, a value of the same order as those obtained in the case of Italy, Switzerland and the 118 English towns. Here we do confront a case in which the nature of the inferences to be drawn depends upon the method of calculation. We may add that the somewhat elaborate calculations have been rather carefully revised and, although errors easily occur, we have some confidence that the results are arithmetically correct. The position is simple, if we trust the method of rates we shall conclude that these English counties do not exhibit a marked correlation between the diseases; if we trust the other method we shall conclude that such correlation is indeed very marked.

We do not propose to re-discuss the whole question of these two methods, which we have examined elsewhere, but we may direct attention to one point. In the tables we have collected correlations based upon various groupings of the counties using the two methods. It will be observed that the correlations based on absolute magnitudes are extremely sensitive to the inclusion or exclusion of certain counties, passing from a very high positive correlation to an inappreciable negative value. On the other hand, the correlations based on rates are much more steady; that is to say that the addition of nine urban counties to the 32 rural and semi-rural counties produces a vastly greater effect in the former than in the latter case. It appears to us that the ultimate reason of this is that, when we employ absolute numbers we

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

weight for size; a town or county with a million inhabitants affects the result much more than does a town or county with 100,000 inhabitants. If, as we consider to be the case, there is no reason why this preferential treatment should be accorded to the more populous districts (we are now assuming that the smallest population in our series is sufficient to justify the supposition that the rate based upon it is materially correct) it appears to follow that the method of rates is the more satisfactory instrument of research and that when the two methods point to different conclusions that inference which is warranted by the method of rates should be accepted.

We are, however, perfectly conscious that this conclusion may not commend itself to all statisticians (for instance, in successive numbers of the *Journal of the Royal Statistical Society*, Professor Karl Pearson and Mr G. Udny Yule have enunciated quite irreconcileable conclusions upon the interpretation of correlations between rates[1]) and we desire to found our present conclusions so far as we can upon analytical results which do not depend upon the acceptance of any disputed or disputable point in the theory of statistics. If, then, we simply confine ourselves to the method of absolute numbers it is singular that we get so great a difference between the correlation in the case of rural and semi-rural on the one side and all the counties, including urban counties, on the other. It will be said at once that this is due to the employment of heterogeneous series, but we have some difficulty in accepting this view. It has been impossible to examine the point properly, because when we try to form still more homogeneous groups and calculate the correlations for purely urban and for purely rural counties separately, we are reduced to 9 and 13 observations respectively.

Calculating the correlation between corrected deaths (numbers) for population constant, we find in the case of the nine urban counties $-\cdot 1358$, and for the thirteen rural counties $+\cdot 4039$, these are both smaller than the coefficient based on the mixed 32, although it is doubtful whether the difference is significant.

If we take the standard deviation of the coefficients to be given by $\frac{1-r^2}{(n-1)^{\frac{1}{2}}}(1+11r^2/4n)$, and use this for determining the standard deviation of differences, it is found that the difference between the value for 32 mixed counties and that for the urban counties is nearly 2·6 times the standard deviation of the difference, while in the case of mixed counties compared with truly rural counties the difference is only 1·5 times the

[1] *J. Roy. Stat. Soc.* **1910, LXXIII. 534–9 and 644–7.**

standard deviation. The former result is perhaps sufficient to render it improbable that the difference can be due to errors of sampling, but the latter is indecisive. It must, however, be remarked that, if the method based on absolute numbers be considered reliable it is significant that the more homogeneous groups (it is evident that the true rural counties are much more alike than are the 32 which grouped together give the highest correlation) yield lower correlations. To put this a little differently, we hold that the true state of affairs as regards the correlation between the two diseases we are studying ought to emerge when we analyse material as homogeneous as possible. In the case of registration counties this ideal is most closely approximated to if we confine ourselves to the rural counties. But in this case we do not in fact obtain so high a correlation as when the material is less homogeneous. Hence one is inclined to argue that the latter result may simply be due to the heterogeneity of the material. This argument is not, however, decisive because the error of sampling which arises when the data are reduced to 13 separate observations is very large and the error distribution imperfectly known so that the customary tests are inapplicable. On the other hand, it is suggestive that the coefficient deduced is of the same order as those obtained by similar methods from 118 towns, the latter being not so homogeneous as the rural counties but probably much more so than the 32 mixed areas.

All the previous reasoning is based on the assumption that the method of using absolute numbers is the correct procedure. If we take instead the coefficient based upon rates, the English results do not differ markedly from those yielded by both methods in the cases of Switzerland and Italy; there is some reason to think that the correlations are a little higher but they are nearer the continental results than the American ones. We are sure that this part of our work is that likely to give rise to the most dispute as to its interpretation, but we believe a careful consideration of the different lines of evidence will incline the reader to conclude that the correlation between the death-rates from cancer and diabetes in the case of English data is, if real, decidedly less marked than was found to be the case in American cities.

General Consideration of the Results.

We are now in a position to discuss the general bearing of our work upon the fundamental problem stated at the outset.

In his note upon Maynard's results, Pearson wrote:

https://doi.org/10.1017/S0022172400005702 Published online by Cambridge University Press

"I think therefore that the relation indicated by Dr Maynard between cancer and diabetes is a real association. It has been here discussed statistically, but no doubt it will be found eventually to have a physiological or pathological basis."

If the passage quoted means simply that Maynard's results cannot be referred to any methodological error, we are in complete agreement with it; both Pearson's analysis and our own seem to prove that Maynard's statistical method was appropriate and sufficient. We may indeed go further than this. Pearson did not specially consider the possibility of material errors; our investigation of this point has not been so complete as we could have wished, but, so far as it goes, it seems to indicate that the sources of error which suggest themselves are quite powerless to account for the coefficients Maynard obtained.

It is, however, possible to read a wider significance into the words quoted. Can it be said that we have statistical demonstration that the association between these two diseases depends upon some physiological identity either of structure or function which results in the production of cancer or diabetes? That is to say that, given two individuals, *A* and *B*, they are, in so far similar and in so far dissimilar that, given the same set of external stimuli, one will develop diabetes and the other cancer and that further, when the proportion of *A*'s in a community is high the proportion of *B*'s will also be high and conversely.

We can interpret this very general statement in a variety of ways:

(1) We might suppose that the peculiarities of *A*'s and *B*'s are inherent in their structure; that they have a physiological or pathological basis. If this were so we should expect to find that the association between the two rates was a universal phenomenon, since, by hypothesis, it depends upon structural peculiarities—part and parcel of the organisation.

It appears to us that our work is cogent, perhaps convincing, evidence against the truth of any such belief. Whatever may be said about certain details, we have not elicited any significant correlation between the disease in the cases of Italy and Switzerland and we have discovered, at the most, a moderate degree of correlation in the case of England. To explain away these results it is necessary to suppose either that (1) all our data—except the 32 mixed counties—are hopelessly vitiated by material or classificatory errors, or (2) that the methods of reduction which were valid in the case of American cities are not applicable to data obtained in Europe. We find it difficult to entertain either hypothesis and, consequently, do not think that there can be any general physiological or pathological basis for the correlation between the two rates as found in America. This conclusion is an important one and we have been at some pains to set out the various reasons which have led us to adopt it. It is not necessary to repeat the evidence again at this place. We are, however, still faced with the fact that we have been unable to offer a satisfactory explanation of Maynard's results. His general hypothesis that the correlation is to be explained as a consequence of the conditions of modern life can, if we are correct, only be true if we suppose that the pressure of competition and other conditions attendant upon life in urban communities are markedly different in America from what prevails in Europe. This is a supposition which seems a little difficult to accept. The other important possibility is that associated with the racial mixture which, as we have proved, so definitely characterises the American cities. Our attempts to gauge the effect of this have been detailed and we have shown that the data are not adequate to permit of a really satisfactory investigation. If our analysis of the other possibly operative factors be regarded as exhaustive, we come by a process of exclusion to this last cause group. Other factors, of which we have no inkling, may play a part; we cannot go beyond our evidence and can merely record the conclusions to which a somewhat laborious analysis has led us.

These conclusions are:—

(1) The death-rates from cancer and diabetes are not universally correlated and the failure to obtain a significant degree of correlation in the case of European statistics does not appear to be due either to errors of method or insufficiency of material.

(2) In the case of England and Wales a significant degree of correlation appears to exist, but its intensity is probably very much less than that observed in the case of American cities.

(3) The correlations obtained in the American cities cannot be explained as results of inexact or inadequate statistical methods nor do they seem to depend upon material errors. The influence of racial heterogeneity may be of considerable importance, but it has been impossible to obtain satisfactory statistical proof of this.

In conclusion we have to express our thanks to Mr J. W. Brown, Assistant in the Statistical Department of the Lister Institute, for much assistance in the arithmetical work of this paper.