SICKNESS, CHANGE OF RESIDENCE, AND DEATH*

I. GENERAL RESULTS OF FOLLOW-UP IN TWO POPULATION GROUPS

BY

DONOVAN J. THOMPSON AND ANTONIO CIOCCO

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

INTRODUCTION

In this paper we present some basic data on residence changes and mortality experience observed in a population sample of (a) the Arsenal Health District of Pittsburgh, Pa, from 1951 to 1956; and (b) Donora, Pa, from 1948 to 1957.

The analyses of these data seek to answer only two questions at this time:

(1) Are the movement (*i.e.* change of residence) and mortality experiences of persons who complain of some sickness in a household interview survey different from the experiences of those who have no such complaint?

(2) To what extent are the estimates of mortality rates affected by the loss of a portion of the originally designated sample through change of address?

Both questions bear upon the increasing efforts being made to describe the natural history of the diseases which now predominate as causes of disability and death, and to identify the significant time periods in which to search for possible aetiological factors of these conditions.

For these purposes studies are needed which can cover a considerable span of time before the onset of disability or death from the chronic diseases. Data obtained through household sickness surveys, of which many have been made in recent decades (Logan and Brooke, 1957), can provide a starting point for

such long-term studies, as we hope to show in this and following papers. This point was illustrated in the resurvey of the Hagerstown population after 20 years (Ciocco, 1946; Lawrence, 1948), which led to the conclusion that "A substantial proportion of the persons who are ill at 65 have had the same or another chronic illness for at least 20 years" (Ciocco and Lawrence, 1952). However, as has been learned from these and many other studies which seek to relate certain individual characteristics at one point in time to the development of disease conditions later, a number of problems of analysis and interpretation arise. These problems are due in part to the bias resulting from the selection of the population under observation (Berkson, 1955), to effects of the passage of time on the development of the condition under study (Kodlin and Thompson, 1958), to the non-specific nature of the characteristic chosen for the initial observation, or to poor definition of the onset of the conditions under study.

In brief, while it is well recognized that long-term studies are required to elucidate the factors related to the onset of the numerically important disabling and fatal diseases of current preoccupation, more knowledge is required to provide a sound basis for carrying out such studies. Moreover, the tools for the analyses of these studies are still somewhat crude, so that the findings resulting from them cannot be readily interpreted. There is a definite need, therefore, to explore and probe further into the statistical as well as into the observational methodology of long-term investigation of chronic disease. The study here reported with others planned in this series constitutes an example of such exploration.

* This work was supported in part by a research contract with the Air Pollution Medical Program, Public Health Service, U.S. Department of Health, Education, and Welfare.

MATERIAL AND METHOD

SOURCE.—The two population groups under study are:

(1) *A sample of the population in a section of the Arsenal Health District of Pittsburgh.*—From this section which, according to the 1950 census, had a population of 81,785 persons, an equal probability sample of 2,954 dwelling units was selected for a household interview survey during July, 1951 (Horvitz, 1952). The survey aimed at obtaining data on such demographic characteristics as age, sex, marital status, nativity, education, etc; on illnesses which had occurred in the month before the survey and services of physicians, hospitals, and clinics received during this period; on hospitalizations and accidents or injuries requiring hospitalization or physician services during the year preceding the survey. Completed interviews were actually obtained from 2,791 households including 9,749 persons, of whom 6,305 were 21 years of age and over in 1951. It is this number that constitutes the base population group followed to July, 1956.

(2) *A systematic one-third sample of the households of Donora, Pa, listed in the 1947 files of the Tax Assessor of the Borough.*—This survey was conducted in 1948 by the U.S. Public Health Service (Schrenk, Heiman, Clayton, Gafaver, and Wexler, 1949). Data were obtained from a responsible person in the household on all members of the household with regard to:

(a) such demographic characteristics as age, sex, marital status, occupation, residence;

(b) past experience of specific conditions such as asthma, sinusitis, heart disease, chronic bronchitis, tuberculosis;

(c) the occurrence of symptoms such as cough, nausea, vomiting, headache, and smarting of the eyes during the smog episode of late October, 1948;

(d) the persistence of these symptoms,

(e) physician and hospital services obtained for the ailments.

Information was obtained by this survey on 4,092 residents of Donora, of whom 2,712 were 21 years of age and over at the time of the survey and constitute that part of the population group followed to 1957 which will be discussed here.

The two population groups differ in many respects as has been indicated and can be surmised; the objectives of the two surveys differed, as did the approach to the selected households and the techniques of obtaining information on illness. Finally, the follow-up mechanism and information was quite different for the two groups.

FOLLOW-UP PROCEDURES.—For the Arsenal sample, subsequent information on change in residence results from a combination of several procedures:

(1) In July, 1952, one year after the original survey, the total sample of households was resurveyed by the same household interview method; and information of the same kind as in the first survey, but more extensive for selected chronic diseases, was obtained. The second survey thus provided data on the number of persons who had moved or died in the year's interval. No attempt was made to trace the persons who had moved. It should be noted also that at the second survey, the interviewers were unaware of the status or other characteristics of the households interviewed at the first survey.

(2) In July, 1956, in connexion with other investigations, the National Cancer Institute requested the Post Office Department for information on changes of address of the whole group of 6,305 persons. The information obtained and made available to us indicated for each individual whether he resided in July, 1956, at the address of 1951, had moved from that address, or was known to have died while residing at that address.

(3) A listing of all deaths of residents of Allegheny County from July, 1951, to July, 1956, was made available to us by the Department of Vital Statistics of the Commonwealth of Pennsylvania. The names of all the 6,305 persons were checked against this listing.

Procedure (3) has provided the basic information on mortality for this study. It revealed, as was to be expected, that the Post Office inquiry missed some of the deaths (approximately one-third) listed in the Vital Statistics Office. Procedure (2) has provided the basic information on the number of persons residing in 1956 at the address of 1951.

As a result of the procedures employed, it can be assumed that we possess fairly reliable information on the identity of the persons who in July, 1956, resided at the 1951 address, and of the persons who died within Allegheny County during the interval 1951–1956. No information is available on the dates of change of residence after 1952 or on the deaths which occurred outside Allegheny County. Supplementary data were obtained through special studies conducted on segments of this population for other purposes. Sub-samples were selected for studies on the prevalence of arthritis (Cobb, Thompson, Rosenbaum, Warren, and Merchant, 1956) and on the prevalence of heart disease (Tauber and Thompson, 1957). The survivors of the arthritis

study group, which to begin with involved 944 persons, are still under periodic observation. The heart study group involved, to start with, 778 persons, who were kept under observation until the middle of 1954. Even with these supplementary data it is anticipated that the identification of deaths and movement is subject to some errors. Specifically, we believe that deaths and movements may be underestimated.

The follow-up procedures in Donora were direct, thus avoiding some of the "errors" inherent in the Arsenal material. These procedures consisted of the following steps:

(1) A complete census of all the households of the Borough of Donora was made in February-March, 1957.

(2) If no member of the household sampled in 1948 remained at the original address, information was sought from the neighbours or current residents as to the present whereabouts of the listed individuals of that household, and, in addition, a central file was established so that information became immediately available if members of that household has been contacted in some other part of the borough.

(3) Letters were sent to every person who had left Donora. As a result, contact was made with that person or with other members of his household in the case of 99.1 per cent. of the persons listed in the original survey.

In the Donora follow-up, information was obtained on the incidence and prevalence of certain chronic conditions during the interval 1948-1957 as well as on hospitalization and medical services received in the interval, place and cause of death, etc. It should be emphasized that the 1957 interviewers were not aware of the responses to health questions which had been made in 1948. Findings related to these aspects of the Donora follow-up will be presented in other publications.

Definitions.—In the 1951 survey of the Arsenal Health District, the respondent was asked questions such as: "Has anyone in the household been ill in bed, been ill at home, or called in or visited a physician during the past month? Has anyone been injured or involved in an accident for which medical care was sought during the past year?"

If the answer to any of these questions was "yes" the person involved was classified as "ill", unless the

reason given for the answer indicated that the episode concerned normal pregnancy or childbirth, routine physical examinations, immunizations, etc.

In the 1948 survey of Donora, the main question was: "Were you affected by the smog of October 28-31?"

If the answer was "yes", the persons involved were placed in the class: "acute illness". Inquiry was also made regarding the prevalence of certain chronic illnesses such as asthma, heart disease, chronic bronchitis, tuberculosis, and sinusitis. Persons with these conditions and/or acute illness are included in the "all illness" class.

One other difference in the study of the two samples is in the definition of movement. For the Arsenal sample, information is available primarily on whether or not the subject was living in 1956 at the 1951 address. More complete information was obtained on the Donora sample. For the purpose of this study, the term "movement" for the Arsenal sample means change in first address, for the Donora sample it means leaving the borough.

DESCRIPTION OF MOVEMENT OR DEATH

In 1956, information was obtained that 368 of the 6,305 persons in the Arsenal sample had died between 1951 and 1956 and that 2,347 had moved from the 1951 address. For all but fourteen of the deaths, copies or abstracts of the death certificates are at hand. For the Donora sample, the resurvey of 1957 found that 298 of the 2,712 had died and that 456 had left Donora in the $8\frac{1}{2}$ years that had elapsed between the two surveys. Copies or abstracts of the death certificates are available on 276 of the dead.

The findings on movement and death are summarized by sex, by broad age groups, and by illness experience reported at the initial survey in Tables IA and IB.

The major points to note are:

(1) The expected difference between the two sexes with respect to mortality is evident, but there is little sex difference with respect to movement.

(2) The percentage of known deaths increase with increase in age at first observation, but the percentages of persons who moved decrease with age.

(3) In the Donora sample (Table IA), for each sex and for each age group, the percentage of ill persons who had died in the $8\frac{1}{2}$ -year interval is larger than the percentage of persons who died during the $8\frac{1}{2}$ -year interval but were not reported ill during

TABLE IA
DONORA SAMPLE, STATUS IN 1957 ACCORDING TO ILLNESS EXPERIENCE REPORTED IN 1948

Age (yrs)	Males					Females				
	In 1948		Percentage in 1957			In 1948	Percentage in 1957			
	Illness* Reported	Number	In Donora	Left Donora	Dead	Number	In Donora	Left Donora	Dead	
21-34	None	238	72.6	26.6	.8	252	69.3	29.9	.8	
	Acute Only	157	64.3	32.5	3.2	203	73.9	24.6	1.5	
	All	187	64.9	32.4	2.7	240	73.9	24.9	1.2	
35-49	None	166	83.7	13.3	3.0	162	85.8	11.7	2.5	
	Acute Only	222	82.4	12.2	5.4	199	81.4	12.1	6.5	
	All	243	82.3	12.3	5.3	232	81.0	12.5	6.5	
50-64	None	136	69.9	14.7	15.4	128	82.8	6.3	10.9	
	Acute Only	235	72.3	8.9	18.7	174	72.4	10.9	16.7	
	All	266	70.7	9.0	20.3	193	72.0	10.4	17.6	
65 and Over	None	44	50.0	4.5	45.5	43	55.8	2.3	41.9	
	Acute Only	82	42.7	7.3	50.0	69	40.6	15.9	43.5	
	All	100	40.0	9.0	51.0	82	39.0	15.9	45.1	

* The "all" illness category includes the "acute only" plus persons reporting only chronic illness.

TABLE IB
ARSENAL SAMPLE, STATUS IN 1956 ACCORDING TO ILLNESS EXPERIENCE REPORTED IN 1951

Age (yrs)	Males					Females				
	In 1951		Percentage in 1956			In 1951	Percentage in 1956			
	Illness Reported	Number	At same Address	Changed Address	Dead	Number	At same Address	Changed Address	Dead	
21-34	No	846	44.9	53.8	1.3	901	47.3	52.7	—	
	Yes	96	54.1	43.8	2.1	161	43.5	54.7	1.9	
35-49	No	862	64.1	33.9	2.0	871	66.5	32.4	1.1	
	Yes	133	51.9	43.6	4.5	213	57.3	36.6	6.1	
50-64	No	597	65.5	25.1	9.4	559	65.5	29.0	5.5	
	Yes	133	54.1	24.8	21.1	187	63.6	28.9	7.5	
65 and Over	No	247	51.8	25.1	23.1	261	56.7	22.6	20.7	
	Yes	100	43.0	18.0	39.0	138	52.1	28.3	19.6	

the smog episode. The differences between the percentages do not increase or decrease consistently from the younger to the older age groups; however, there is a tendency for the ratios of the corresponding percentages to decrease from the younger to the older age groups.

(4) The percentage of ill persons who left Donora in the $8\frac{1}{2}$ -year interval is higher than the corresponding percentage among persons not reported ill who left Donora in two of the four age groups among males and three of the four age groups among females. Thus it would appear that there is no consistent difference between the ill persons and the persons not reported ill with respect to migration from Donora.

(5) In the Arsenal sample (Table IB) for each sex and for each age group (except females 65 years of age and over) the percentage of persons who died

during the interval 1951-1956 is higher among the persons reported ill in the 1951 survey than among those not reported ill. This difference is more striking in the males than in the females.

(6) Among the females in three of the four age groups in the Arsenal sample, the percentage of persons who moved is higher for those who reported illness in the 1951 survey than those who did not. Among the males this is observed in only one age group.

Before proceeding to an estimate of the mortality rates, let us consider the above findings. Taken at face value they indicate that reported illness, be it the acute variety associated with a smog episode or a conglomeration of acute and chronic conditions as found in the Arsenal survey, segregates from the population a group which for a few years at least will have a higher mortality than the rest. Further

analyses of our data will seek to determine the circumstances in which the differences persist. Some illnesses or conditions are undoubtedly more important than others in such segregation. From the data of the Hagerstown resurvey which took place after an interval of 20 years, the conclusion was reached that the life span was shortened only when major chronic conditions were reported in the initial survey. Future examination of the material at hand on the Arsenal and Donora samples will indicate how valid such conclusion is for the shorter span of time involved here. The lack of striking differences in mortality between the ill and not-ill females in the oldest age group of the Arsenal sample also deserves additional examination, since it may reflect the respondent bias often suspected in household surveys.

In a subsequent paper it will be shown that the differences in movement rates between the ill and not-ill persons probably reflect only sampling variation. Nevertheless, the amount of the change is considerable, and is an important factor to consider in any attempt at follow-up of population samples. It creates an important problem for the estimation of mortality rates as we shall next point out.

ESTIMATION OF MORTALITY RATES

The data we have described relate mortality and mobility through time in two groups of persons—the “ill” and the “not ill”. In many follow-up investigations, the persons we have described as “moved” would be lost to follow-up on the date of their first movement. In some investigations diligent efforts are made to keep this moved group under observation, but nevertheless some fraction (large or small) is finally lost to follow-up. The careful investigator is concerned with the possible influences such losses may have in contrasting two or more groups, such as our “ill” and “not ill”, in terms of mortality or any other characteristic. From this point of view the movement experience can be viewed as a disturbance the possible influence of which requires investigation. The principal fear appears to be that the remnants remaining under observation at the end of some time period will not be “representative” of their respective initial groups. Differential selection of the “movers” in the various groups is, perhaps, viewed as the principal mechanism responsible for any lack of comparability among the remnants of each group at the end of some time period.

In this material, the data for the Donora sample include information about the date of movement from the borough of all the persons who moved, and about the date and place of death of those who

died during the interval 1948 to 1957. For the Arsenal sample, in the interval 1952–56, information is available only about the date and place of death of those who died and could be found in the vital statistics records, but no information was obtained about the date of change of residence within the period. Since the Donora data are more complete, the analytical procedures to be described will be based on these data to begin with.

The events of death and movement in a series of discrete intervals of time may be shown graphically as in the Figure (opposite). During the first interval, (t_0, t_1) , the persons in the cohort have a probability of dying, P_1 , and a probability of moving, R_1 . By the end of the first period, t_1 , the persons alive at the beginning time point, t_0 , will be found in one of the four categories indicated by the four following cells:

Category	Probability
Moved and Died	$\phi_{12}(1)$
Moved and Survived	$\phi_2(1)$
Not moved and Died	$\phi_1(1)$
Not moved and Survived	$\phi_0(1)$

In the next interval, t_1 to t_2 , we have two categories of living persons to consider (indicated by the broken arrows): those who moved and those who did not move during the first period. If we are not interested in the subsequent movement of those who moved in the interval t_0 to t_1 , we can consider that during the interval t_1 to t_2 they are subject only to the risk of dying, the probability of which is designated as $P_2^{(1)}$. The survivors of the first period who did not move are again subject to both movement and death during the interval t_1 to t_2 , and their probability of dying in the interval is designated as P_2 , which may be the same as $P_2^{(1)}$ or may have a different value. The survivors at t_2 are of three types, as may be noted from the Figure, and for the period t_2 to t_3 a probability of dying is defined for each of the three types, P_3 , $P_3^{(2)}$, $P_3^{(1)}$, which may again be identical or different.

In terms of this representation of the events of movement and death and their associated probabilities, the total probability of dying during the interval t_1 to t_w , designated P_t , is given by:

$$P_t = \left[\sum_{i=1}^w \frac{\pi^{i-1}}{\pi} (1 - P_i) \right]^{-1} \sum_{j=1}^w \left\{ \frac{\pi^{j-1}}{\pi} \phi_0(j) \phi_2(j) \frac{\pi^{w-1}}{\pi} (1 - P_i) \left[1 - \frac{\pi^j}{\pi} (1 - P_i) \frac{\pi^w}{\pi} (1 - P_i^{(j)}) \right] \right\}.$$

The observed frequency of occurrence of the various categories of events in the model will furnish the basic information for estimating the probabilities defined. For the case of complete follow-up for the period t_0 to t_w , P_t would be estimated simply by dividing the total observed deaths (ignoring movement) by the number of persons in the original sample.

If the follow-up of the movers is incomplete, as is most frequently the case, we may be forced to base our calculations of mortality rates on the incomplete experience obtained by following each individual to the end of the period during which he moves. From the model described in the Figure we shall then be estimating a total probability of dying during the interval t_0 to t_w defined as P_s , where

$$P_s = 1 - \prod_{i=1}^w (1 - P_i).$$

FIGURE.—Schematic representation of a probability model for follow-up investigations of a characteristic (Mortality- d) in the presence of a disturbance (Movement- m). Three consecutive time periods represented.

In this equation, k is a proportionality constant and P_m^* has the same functional form as P_m , with $P_i^{(i)}$ replaced by P_i throughout.

Further to clarify the interpretation of P_m , we might calculate the survivors to say, 1953, of the sub-group which moved in 1948 and survived that period, and the survivors of the subgroups which moved in the years 1949–1952. The total number of persons in these subgroups multiplied by $(1 - P_m)$ will give the total survivors to 1953 of these subgroups. It can also be shown that this calculation yields the same result as would be obtained by employing the usual actuarial or life-table methods appropriate for open cohorts.

For the Donora data, the probabilities so calculated are shown in Table II, which shows:

(1) The probability of dying (expressed by P_i) is greater among the ill than among the not ill for all ages and both sexes. However, the differences for corresponding intervals is not constant but increases to about 1948–53 and decreases thereafter.

(2) In the younger age group, the probability of dying for those who moved (as expressed by P_m) is zero, but in the older age group it is, for some intervals, slightly higher than the total probability (P_i) or the probability of dying while in Donora (P_s). The numbers involved during the intervals when this is observed are too small to give reliable rates. For the same reason not too much weight can be given to the higher values of P_m in the comparison of the ill and the not ill.

It must be recalled that the entrance into the moved group of a subgroup with subsequent low mortality will produce a decrease in the value P_m at that point, as is seen in Table II for ages 50 and over for the cumulative periods 1948–53.

(3) In the younger age group, the values of P_s are uniformly equal to or greater than the corresponding values of P_i . In the older age group, the influence of mortality among those who moved is reflected in the fact that for some intervals P_s may be smaller than P_i . For the whole period 1948–1957, P_s is greater than P_i for both age groups and both sexes.

(4) The relative differences, $(P_s - P_i)/P_i$, range from zero to as much as 18 per cent., the numerically larger relative differences being associated in general with periods for which the rates are low.

These findings bring out two important points for consideration. First, that the influence of the factor of the one episode of illness on subsequent mortality would become undetectable in the long run if the apparent trend continued. One may conclude from from this that, in long-term studies in which the relationship is sought between the manifestation of a condition and subsequent mortality or other measure of health, there will be an optimal period in which it is possible to measure the influence of the factor. Secondly, the mortality rates among those who moved and those who stayed are different. This strengthens the idea that a certain risk is involved in determining mortality rates on the assumption that

TABLE II

CUMULATIVE PROBABILITY OF DYING BETWEEN 1948 AND SUBSEQUENT YEARS IN DONORA SAMPLE, ACCORDING TO SEX, AGE, AND ILLNESS STATUS IN 1948, AND MOVEMENT STATUS IN SUBSEQUENT YEARS

Sex	Time Interval	Age (yrs)											
		21–49						50 and Over					
		Acute Illness in 1948			No Illness in 1948			Acute Illness in 1948			No Illness in 1948		
		P_i	P_s	P_m	P_i	P_s	P_m	P_i	P_s	P_m	P_i	P_s	P_m
Male	1948–49	0.005	0.005	—	0.002	0.002	—	0.022	0.022	—	0.022	0.022	—
	1948–50	0.011	0.011	0.0	0.002	0.002	0.0	0.057	0.057	0.0	0.039	0.039	0.0
	1948–51	0.018	0.019	0.0	0.007	0.008	0.0	0.114	0.116	0.0	0.061	0.051	0.286
	1948–52	0.021	0.022	0.0	0.007	0.008	0.0	0.142	0.145	0.0	0.100	0.091	0.286
	1948–53	0.021	0.022	0.0	0.012	0.013	0.0	0.180	0.187	0.0	0.128	0.121	0.182
	1948–54	0.032	0.035	0.0	0.012	0.013	0.0	0.199	0.208	0.053	0.156	0.151	0.154
	1948–55	0.034	0.038	0.0	0.017	0.019	0.0	0.246	0.255	0.053	0.189	0.188	0.125
	1948–57	0.045	0.051	0.0	0.017	0.019	0.0	0.268	0.279	0.045	0.228	0.233	0.095
Female	1948–49	0.002	0.002	—	0.002	0.002	—	0.021	0.021	—	0.012	0.012	—
	1948–50	0.015	0.015	0.0	0.002	0.002	0.0	0.058	0.058	0.0	0.018	0.018	0.0
	1948–51	0.022	0.023	0.0	0.002	0.002	0.0	0.078	0.079	0.0	0.035	0.035	0.0
	1948–52	0.022	0.023	0.0	0.005	0.005	0.0	0.111	0.113	0.0	0.053	0.053	0.0
	1948–53	0.025	0.026	0.0	0.005	0.005	0.0	0.144	0.148	0.0	0.082	0.078	0.250
	1948–54	0.032	0.035	0.0	0.005	0.005	0.0	0.189	0.197	0.0	0.117	0.114	0.167
	1948–55	0.032	0.035	0.0	0.012	0.014	0.0	0.222	0.228	0.056	0.152	0.151	0.167
	1948–57	0.040	0.044	0.0	0.014	0.017	0.0	0.243	0.247	0.091	0.187	0.188	0.111

there is equality in the mortality rates of those who moved and those who did not.

The P_t as calculated is a measure of the probability of dying of the total cohort. Unless it is possible to follow-up a cohort, as was done in Donora, this probability can be estimated only approximately. It will be noted that, if P_s were used in its stead, the value of P_t would be overestimated when the mortality of the moved was less than the mortality of those who did not move, and that the value of P_t would be underestimated when the mortality of the moved was higher than that of those who did not move.

Several ways have been suggested for estimating the probability of dying when complete information on the persons who have moved is not available. For example, there is the time-honoured approach used in life-table construction in which the population at risk at the beginning of the interval is reduced by one-half of the persons who moved during the interval. With this approach,

$$\hat{P}_i = \frac{d_i}{1_i - \frac{1}{2}m_i},$$

when d_i = observed deaths and m_i = the number moved (lost to follow-up) during the period, and 1_i = the number at risk at the beginning of the period. For w periods of follow-up we may define P_a as follows:

$$P_a = 1 - \pi^w (1 - \hat{P}_i).$$

If, as is often the case, information is not available regarding the dates of movement and subsequent mortality, and if there is only information on the number of persons at the initial observation period, the numbers who moved, and the numbers who died in the place of residence, only an over-all mortality rate for the total interval of follow-up can be calculated by following the suggested approach. The rates calculated in this manner (P_s , P_a) may be denoted, respectively;

$$P_s^\circ = \frac{d}{d + s}$$

$$P_a^\circ = \frac{d}{n - \frac{1}{2}m}$$

where: n is the size of the original cohort

d is the total number of observed deaths

s is the number of survivors among those who have not moved,

m is the number of persons who have moved (lost to follow-up) during the total period of follow-up.

The values of P_s , P_a , P_s° , P_a° , calculated on the 1948-1957 experience of the Donora sample, are compared with P_t in Table III; and the values of P_s and P_a calculated on the one year experience (1951-1952) of the Arsenal sample are compared with P_t in Table IV.

TABLE III

COMPARISON OF VARIOUS ESTIMATES OF THE PROBABILITY OF DYING, DONORA SAMPLE, 1948-1957, ACCORDING TO SEX, AGE, AND ILLNESS STATUS IN 1948

Probability of Dying	Male				Female			
	Age 21-49		Age 50 and Over		Age 21-49		Age 50 and Over	
	Ill	Not Ill	Ill	Not Ill	Ill	Not Ill	Ill	Not Ill
P_t	0.0449	0.0173	0.2681	0.2278	0.0398	0.0145	0.2428	0.1871
P_s	0.0506	0.0193	0.2791	0.2327	0.0438	0.0173	0.2467	0.1878
P_a	0.0513	0.0196	0.2784	0.2351	0.0445	0.0175	0.2488	0.1829
P_s°	0.0500	0.0194	0.2772	0.2321	0.0438	0.0164	0.2511	0.1813
P_a°	0.0565	0.0220	0.2907	0.2500	0.0488	0.0188	0.2701	0.1875

TABLE IV

COMPARISON OF VARIOUS ESTIMATES OF THE PROBABILITY OF DYING, ARSENAL SAMPLE, 1951-1952 AND 1951-1956, ACCORDING TO SEX, AGE, AND ILLNESS STATUS IN 1951

Period	Probability of Dying	Male				Female			
		Age 21-49		Age 50 and Over		Age 21-49		Age 50 and Over	
		Ill	Not Ill	Ill	Not Ill	Ill	Not Ill	Ill	Not Ill
1951-2	P_t	0.0089	0.0059	0.0772	0.0178	0.0080	0.0006	0.0277	0.0244
	P_s	0.0103	0.0069	0.0794	0.0189	0.0000	0.0000	0.0270	0.0209
	P_a	0.0094	0.0063	0.0761	0.0183	0.0000	0.0000	0.0258	0.0202
1951-6	P_s°	0.0337	0.0150	0.2687	0.1187	0.0279	0.0022	0.1064	0.0806
	P_a°	0.0472	0.0210	0.3195	0.1422	0.0400	0.0030	0.1318	0.0982

The data of Table III reveal immediately that, in comparison with P_t , the "true" mortality rate for the 8½-year period, the other calculated rates are higher in most instances. The estimate, P°_s , has the largest discrepancy, obviously because it entirely overlooks the moved in the denominator. On the other hand, when the rates are calculated annually, based only on the persons who stayed in Donora (P_s), the magnitude of the deviation from P_t is about equal to if not smaller than that observed when P_a or P°_a is computed. It is interesting to note that the difference between P°_a and P_t is more often smaller than the difference between P_a and P_t . Similar comparisons are possible on the Arsenal sample for the one-year interval 1951–52.

The data of Table IV show that in this instance the differences between P_a and P_t are more often smaller than the difference between P_s and P_t .

The implication of these comparisons will be discussed elsewhere. Here it is sufficient to point out that the differences between the several rates compared are in part due to differences in the frequency of movement within the time intervals studied. Consequently, estimation procedures for mortality or other rates become purely arithmetical exercises unless they take into account the actual behaviour of disturbing factors such as movement in the above example.

It is of interest to note from Tables III and IV that the difference between corresponding rates for the "ill" and "not ill" or the ratio of the corresponding rates indicates that the differences or ratios based on P_t are closely approximated by differences or ratios based on the various approximations. In other words, as the events of movement occurred in the Donora Sample for the period 1948–57 and in the Arsenal Sample for the period 1951–2, the approximation usually employed to calculate mortality when there is loss in the population at risk (here indicated as P_a) may overestimate or underestimate the actual mortality rates of the ill and not ill by a considerable amount. However, as it happens, these approximations do not seem to distort too much the difference between the two groups.

SAMPLING VARIATION

Chance fluctuations in all the rates we have described must be considered when contrasting the ill group with the not ill group. To employ a parametric test for this purpose requires calculation of the variance. If we follow accepted procedures, the variance of the estimated P_t may be taken simply as $P_t(1 - P_t)/n$, where n is the original size of the group considered. For rates such as P_a , in which

corrections are made for the reduction in the size of the original cohort, a formula suggested by Irwin (1949) may be used. The difference in mortality rates for corresponding ill and not ill groups, when divided by the estimated variance of the difference, will be approximately normally distributed with zero mean and unit variance under the "null hypothesis".

From the data of Tables III and IV we have carried out tests of the differences between mortality rates for corresponding ill and not ill groups, based on P_t for 1948–57, for the Donora sample, and on P_t for 1951–52, and P°_a for 1951–56, for the Arsenal sample. The results of the tests show that the observed difference would occur by chance less than once in a hundred in the cases of:

(a) Donora males and females in the age group 21–49;

(b) Arsenal males of 50 years and over, when P_t for 1951–52, is compared;

(c) Arsenal males of 50 years and over and females of 21–49 years when P°_a , for 1951–56 is compared.

For the other age and sex groups on which the tests have been made the differences are apparently not large enough to reach statistical significance in terms of the stated criterion for sampling variation.

The whole question of testing group differences for follow-up data of the type reported here is, of course, dependent upon the specific objectives of the analysis. Differences in the cumulative mortality between the two groups up to some point of time in the follow-up period may be of interest for one purpose; differences in the annual mortality may be of interest for another. Perhaps, most generally, we would seek for a representation of the entire phenomenon in terms of the cumulative probability distribution function of the time at death to determine whether the two groups have different distributions and what the magnitudes of the differences are.

To our knowledge, an adequate discussion of the problems of determining the proper variance expressions, the "effective" sample size, the forms of the distribution functions, choice of test criteria, etc., which are appropriate for follow-up investigations has not as yet appeared in the literature. In subsequent papers of this series we hope to examine this aspect of the analysis of cohort data in greater detail.

DISCUSSION

The relationship of reported illness and subsequent mortality and change of residence has been

examined (a) for the interval 1948–1957 in a sample of 2,712 persons aged 21 years and over residing and surveyed in Donora, Pa, in 1948; and (b) for the interval 1951–1956 in a sample of 6,305 persons aged 21 years and over residing and surveyed in the Arsenal Health District of Pittsburgh, Pa, in 1951. The survey of Donora in 1948 was undertaken to investigate the sickness effects of the smog episode of October of that year. The survey of the Arsenal District in 1951 was undertaken to determine the frequency of illness and the receipt of medical care in the population.

The analyses discussed in this paper seek to answer only two questions:

(1) Did persons who reported illness change their address or die at a different rate from persons who did not?

(2) To what extent is the estimate of mortality affected by the differential mobility?

In answer to the first question, the data reveal:

(1) There was no consistent difference in the frequency of change of address between those who reported illness at the initial survey and those who did not.

(2) In general, persons who reported some illness at the initial survey had a higher mortality than those who did not.

(a) For the Donora data for which annual mortality rates could be estimated, the difference showed a tendency to become smaller with length of time.

(b) The difference between the mortality of those who reported illness at the initial survey and those who did not may be due to chance variation for certain age-sex subgroups. However, a thorough investigation of the role of chance in the observed differences has not been undertaken.

While a full interpretation of these findings requires further analysis of the data, certain inferences may be drawn now. It appears that observations such as these made by eliciting information on sickness complaints through interviews does identify persons who for some time will have a higher risk of dying than others. Taking into consideration the Hagerstown study already mentioned and the findings of the Donora sample, we could conclude that the duration of this higher risk has an optimum point.

If it were true that identification of persons with a greater risk of mortality for a specified interval of

time can be made through the interview type of survey, this approach could be used to throw light on the patterns of sickness which lead to death and also to provide an earlier starting point for studies aimed at exploring antecedent factors in sickness. Further analysis of the present material, particularly with respect to the relationship of specific sickness to specific cause of death, and of specific sickness in successive observations, is required before a final interpretation can be made of the findings here reported.

In answer to the second question, the Donora data indicate that estimates in which an adjustment is made for loss of data due to change of residence, or in which this loss is disregarded, may vary considerably from the rate based on knowledge about the total cohort. Although, in this particular sample, the different estimates did not affect the differences in mortality rate between those who reported illness at the initial survey and those who did not, it is clear that the usual methods of estimating mortality rates for a cohort that has been affected in the course of time by a loss of persons under observation may give rather crude approximations. Our data have shown that the differential rate of mortality among those who move as well as the varying proportions of persons who move will produce departures from the actual mortality rate that cannot be compensated for by the adjustment procedures now employed.

The observations and points discussed only hint at some of the major questions which must be resolved before we can undertake with assurance the analysis of long-term or cohort data. It is hoped that as more such data are collected greater efforts will be made to re-examine critically the pertinent analytical procedures.

SUMMARY

Data on reported illness, subsequent mortality, and change of residence from two follow-up studies of approximately 9,000 adult persons have been analysed for relations between reported illness status and subsequent mobility or subsequent mortality. These analyses show (1) no consistent or clear-cut relation between reported illness status and later mobility, and (2) higher subsequent mortality among persons reported ill than among persons reported not ill. Various methods of estimating the subsequent mortality in the ill and not-ill group, which utilize different assumptions concerning mortality within the groups lost to follow-up through movement, are compared with the rate based on complete mortality knowledge about the total cohort.

REFERENCES

- Berkson, J. (1955). *Proc. Mayo Clin.*, 30, 319.
- Ciocco, A. (1946). *Hum. Biol.*, 18, 33.
- and Lawrence, P. S. (1952). In "Illness and Health Services in an Ageing Population", by G. St. U. Perrott, A. Ciocco, G. Baehr, L. S. Rosenfeld, and Their Collaborators, p. 26 (Public Health Service Publication, No. 170). Washington, D.C., U.S.A.
- Cobb, S., Thompson, D. J., Rosenbaum, J., Warren, J. E., and Merchant, W. R. (1956). *J. chron. Dis.*, 3, 134.
- Horvitz, D. G. (1952). *Publ. Hlth Rep. (Wash.)*, 67, 1003.
- Kodlin, D., and Thompson, D. J. (1958). "An Appraisal of the Longitudinal Approach to Studies of Growth and Development". *Monogr. Soc. Res. Child Developm.*, 23, 1. (Serial No. 67).
- Lawrence, P. S. (1948). *Publ. Hlth Rep. (Wash.)*, 63, 69.
- Logan, W. P. D., and Brookes, E. M. (1957). "The Survey of Sickness, 1943-1952". General Register Office. Studies on Medical and Population Subjects, No. 12. H.M.S.O., London.
- Schrenk, H. H., Heimann, H., Clayton, G. D., Gafafer, W. M., and Wexler, H. (1949). "Air Pollution in Donora, Pa. Epidemiology of the Unusual Smog Episode of October, 1948. Preliminary Report" (Public Health Bulletin, No. 306). Washington, D.C., U.S.A.
- Tauber, J., and Thompson, D. J. (1957). *J. chron. Dis.*, 6, 595.