Study of disease associations from linked records

M. S. T. HOBBS, A. S. FAIRBAIRN, E. D. ACHESON, AND J. A. BALDWIN

*Community Medicine and Primary Medical Care, University of Southampton and Unit of Clinical Epidemiology
University of Oxford*

Hobbs, M. S. T., Fairbairn, A. S., Acheson, E. D., and Baldwin, J. A. (1976). *British Journal of Preventive and Social Medicine*, **30**, 141-150. **Study of disease associations from linked records.** Comparisons are made between the principal diagnosis on discharge from hospital in successive periods of inpatient care for persons in the Oxford Record Linkage Study area admitted over a period of years. The observed numbers of pairs of diagnoses are compared with expected numbers computed to take account of the discharge rates in the population by age and sex and the number of man years of exposure. Three topics have been selected to display some of the types of analysis possible with the very extensive material: mental disorder and diseases of the central nervous system, hospital discharges *preceding* those for neoplasm, and discharges *following* those for tuberculosis. The details of the method of calculation of the observed and expected numbers and the assumptions and approximations involved are given in an Appendix.

One of the advantages of a system of linked medical records is that it makes possible the study of the relationship between successive incidents in a patient's medical history. The study described here compares the principal diagnosis on discharge from hospital in successive periods of inpatient care for persons admitted from a defined population over a period of years. The file contained particulars of 193 710 discharges experienced by 138 508 patients. The observed numbers of pairs of diagnoses on successive discharges were compared with expected numbers computed in a manner which took account of the discharge rates in the population by age and sex, and the number of man years of exposure. Although all the material relates, in strict terms, to discharges—that is, to completed spells of inpatient treatment—the terms 'admission' and 'discharge' are used synonymously.

MATERIAL AND METHOD

Coded information on every birth, death, hospital inpatient discharge, and obstetric delivery has been recorded for residents in the Oxford area since 1963 and in the Reading area since 1966; it

is stored on magnetic tape. Through personal identifying information accompanying each record, events occurring in the same individual can be identified and brought together on the tape (linked). The source material for the present study was a linked file of general and psychiatric hospital discharges and of deaths occurring for residents in the Oxford area (average population of 355 000) during the five-year period 1963-67 and for residents in the Reading area (average population of 414 000) for two years 1966-67. Because of technical difficulties, deaths of residents outside hospital were not included for the first two of the five years covered by the study. The following variables were used in the analysis: sex; age in five-year groups; principal diagnosis on discharge in accordance with the 3-digit code of the 7th revision of the *International Classification of Disease (ICD)*; surgical operation, if any, according to the 1956 Operation Code of the General Register Office; death (yes or no), date of death, and area of residence (Oxford or Reading).

The tape was scanned by computer to identify patients who had experienced two or more discharges with differing principal diagnoses. Re-admissions for the same principal diagnosis were

excluded. The number of patients in whom one diagnosis was followed by another diagnosis was counted and accumulated in a giant matrix ($999 \times 999 \times 23$ cells) held within the computer store. The number of times each diagnostic combination would be expected to occur purely by chance was then calculated from the frequency of each diagnosis in the population of the area from which the patients were drawn. The method of computation of the expected numbers, which included a correction for age, is given as an Appendix to this paper together with a discussion of the assumptions and approximations used.

A comprehensive table was printed, showing all combinations of diagnoses incurred by one or more patients, and a long list of all diagnostic combinations in which the expected number of cases exceeded 0.001. These are available for inspection at the Unit of Clinical Epidemiology. A shorter list was also printed by computer showing for a limited number of diagnostic pairs the first diagnosis, the second diagnosis, and the observed and expected number of occasions on which this sequence of diagnoses occurred. In order to minimize the attenuation of the results

by small numbers, the observed and expected numbers comprised all ages and both sexes for both areas. The printed output was further reduced by excluding diagnostic pairs unless the following two criteria were valid. First, that the observed count should be three or more, and secondly that the ratio of the observed to the expected number should be three or greater.

It was thought that statistical tests of significance would not be appropriate. Their calculation would be wasteful of computer time and they would be of doubtful validity because of the imprecise nature of the data. One page of the shorter list, expanded by inclusion of the written diagnoses, is shown in Table I. It should be noted that the 'expected' numbers in Table I (and throughout the remainder of the material) refer to discharges, that is, to *events*, and that the numbers may therefore be freely added together to give an expected number for a group of diagnoses. The 'observed' number in these tables refers to patients, that is, to *persons*, and are, strictly, not additive in the same way since one patient might have readmissions with different diagnoses and so contribute to more than one observed number. However, it proved

TABLE I
PART OF SHORTER LIST OF DIAGNOSTIC COMBINATIONS, SHOWING OBSERVED AND EXPECTED NUMBERS OF CASES
AND RATIO OF OBSERVED TO EXPECTED

Diagnosis 1		Diagnosis 2		Combination		
ICD Code	Condition	ICD Code	Condition	Observed	Expected	O/E
002	Pulmonary tuberculosis	003	Pleural tuberculosis	4	0.3	13.3
		016	Genitourinary tuberculosis	4	0.1	40.0
		309	Other psychoses nec	4	0.5	8.0
		320	Pathological personality	3	0.5	6.0
		410	Diseases mitral valve	3	0.6	5.0
		433	Functional disease heart	3	0.7	4.3
003	Pleural tuberculosis	002	Pulmonary tuberculosis	5	0.3	16.7
		162	Malignant neoplasm bronchus and trachea and lung specified as primary	6	1.7	3.5
		170	Malignant neoplasm breast	3	0.6	5.0
010	Tuberculosis meninges and CNS	002	Pulmonary tuberculosis	4	0.02	200.0
050	Scarlet fever	510	Hypertrophy tonsils and adenoids	5	0.4	12.5
051	Streptococcal sore throat			4	1.2	3.3
056	Whooping cough	475	Acute upper respiratory infection multiple or nec	3	0.5	6.0
082	Acute infectious encephalitis	343	Encephalitis, myelitis, and encephalo-myelitis	3	0.01	300.0
085	Measles	056	Whooping cough	5	0.2	25.0
		087	Chicken pox	3	0.1	30.0
		353	Epilepsy	3	0.2	15.0
		763	Lobar pneumonia	3	0.3	10.0
		785	Symptoms referable to abdomen and lower gastrointestinal tract	4	0.6	6.7
096	Other diseases attributable to viruses	374	Keratitis	4	0.01	400.0
134	Other fungus infections	763	Bronchopneumonia	3	0.3	10.0

technically too laborious to recalculate exactly the observed number of patients readmitted for a group of diagnoses: both observed and expected numbers in this table have been freely added together to give figures for grouped diagnoses in subsequent tables. The possible small error introduced by this procedure, which tends to inflate the numbers, should be borne in mind in the interpretation of results.

RESULTS

Although it was theoretically possible to look at any combination of two 3-digit diagnostic codes in successive discharges experienced by the same patient, the volume of material available for study was so large that we have not tried to cover exhaustively all findings of clinical interest. We have therefore selected three topics in which the results seemed of importance in order to display some of the possible types of analysis: mental disorder and diseases of the central nervous system (CNS), hospital discharges preceding those for neoplasm, and discharges following those for tuberculosis.

MENTAL DISORDERS AND DISEASES OF THE CENTRAL NERVOUS SYSTEM

The main rubrics for mental disorders and

diseases of the central nervous system lie within the range 300-357 in the *ICD* (7th Revision). However, other important conditions or related conditions are to be found outside the main block. These include: malignant and benign cerebral tumours, symptoms referable to the central nervous system and senility, head injuries—including fractures of the skull, concussion, and other intracranial lesions. Poisoning should also be included in any study of readmissions of mental disorders on the assumption that most of such admissions are the result of self-poisoning. In order to explore the relationship between mental disorders and organic neurological disorders, we have combined the individual diagnoses into eight broad groups as follows: mental disorders (300-326); stroke (340-345); epilepsy (353); cerebral tumour (193-223); other specified CNS disorders (340-357); CNS symptoms and senility (780-781), (790-791); head injury (N800-N804, N850, N852-856); and poisoning (N960-979).

In Table II, the first admissions within each group of diagnoses have been tabulated to show the number of readmissions for a diagnosis within any other group. Table II also shows readmissions for any cause within the same group but excludes readmissions for exactly the same 3-digit diagnosis.

TABLE II

READMISSIONS OF PATIENTS WITH MENTAL DISORDERS, DISEASES OF THE CNS, HEAD INJURY, AND POISONINGS

Diagnosis on Second Admission	Diagnosis on first admission																	
	Mental Disorders		Stroke		Epilepsy		Primary Cerebral Tumour		Other Specific Diseases of CNS		Symptoms Related to Nervous System and Senility		Head Injury		Poisoning		All Mental and CNS Disorders	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
Mental disorders	796	16.5	45	34.4	54	3.0	1	1.0	58	17.9	204	9.9	75	20.7	507	10.6	1740	114.0
Stroke	66	34.4	169	35.7	9	1.7	4	0.6	101	38.8	19	15.7	20	11.0	5	5.3	393	143.2
Epilepsy	27	3.0	5	1.7	—	—	2	0.2	9	1.0	16	0.7	20	4.9	12	2.3	91	13.8
Primary cerebral tumour	7	1.0	12	0.6	9	0.2	18	0.0	13	1.0	7	0.2	3	2.0	1	0.9	70	5.9
Other specific diseases of CNS	60	17.9	175	38.8	10	1.0	6	1.0	87	1.0	39	6.1	3	7.3	4	3.6	384	76.7
Symptoms related to nervous system and senility	239	9.9	10	15.7	7	0.7	—	0.2	14	6.1	13	0.4	17	5.2	8	3.7	308	41.9
Head injury	45	20.7	4	11.0	18	4.9	2	2.0	10	7.3	21	5.2	38	7.6	30	9.9	168	68.6
Poisoning	259	10.6	10	5.3	10	2.3	1	0.9	9	3.6	3	3.7	60	9.9	168	36.3	520	72.6
All mental and CNS disorders	1499	114.0	430	143.2	117	13.8	34	5.9	301	76.7	322	41.9	236	68.6	735	72.6	3674	536.7

CNS = central nervous system

For example, in 796 pairs of admissions for mental disorder there was a change in the specific diagnosis within the rubric for mental disorders.

The table shows that in total there were 3674 readmissions with a different diagnosis falling within the category of mental disorder or disease of the central nervous system. Of these, 1289 (35.0%) were for a different diagnosis within the same general group (that is, the sum of the diagonals in the table). It is probable that this was, in most cases, because of diagnostic uncertainty, lack of specificity, or error in coding on one or other occasions rather than an actual change in the disorder for which the patient was treated. As might be expected, by far the largest number of readmissions within a single group (796) was for mental disorder. It was also noted that mental disorder was preceded (507 cases) or succeeded (259 cases) by poisoning. Probably most, if not all, of these were the result of self-poisoning or therapeutic misadventure related in either case to the underlying mental disorder. Within other groups there was a change of diagnosis in 169 patients admitted twice for stroke and in 87 with another specific disease of the central nervous system. In 18 instances there was a change in the specific category of cerebral tumour recorded. In addition to the 1289 readmissions coded to different diagnoses within the same general group, a further 617 readmissions occurred in which either the first or second diagnosis (but not both) was recorded in indefinite or symptomatic terms. Thus, in no fewer than 1906 (51.8%) of the readmissions shown in Table II either diagnostic uncertainty or an error in coding is likely to have played a part in the discrepancy between the first and second diagnoses. Table II thus illustrates the potential of a linked file of records in revealing deficiencies in the classification and coding of hospital morbidity data. This could provide a useful tool for both systematic investigations of present shortcomings in the coding of morbidity data, and for monitoring the quality of information collected in large regional or national surveys of hospital morbidity.

In 205 admissions for mental disorder, there followed a readmission with a specific neurological diagnosis including head injury. This compared with an expected number of 77.0. In detail the diagnosis on readmission was: stroke 66 (expected 34.4), epilepsy 27 (expected 3.0), cerebral tumour 7 (expected 1.0), other diseases of the central nervous system 60 (expected 17.9), and head injury 45 (expected 20.7). A further 259 readmissions were ascribed to poisoning (expected 10.6).

The slightly increased risk of readmission for stroke is not unexpected and is consistent with the

presence of degenerative vascular disease as an underlying cause for both mental disorders and stroke. Similarly, the ninefold increase in risk of readmission for epilepsy after admission for mental disorder can also be explained on the grounds of underlying organic brain disease. The small group of seven cases initially admitted for mental disorder but later admitted for cerebral tumour is of particular interest clinically as an example of possible diagnostic error which will be discussed in further detail in relation to Table III. Finally the group of patients with mental disorders readmitted for head injury is noted to be more than twice the expected number. This apparent predisposition of patients with mental disorder to injury requiring hospitalization may have several explanations. These include abuse of drugs—particularly alcohol, adverse effects of therapeutic drugs given for psychiatric reasons, underlying organic brain disease leading to impaired physical performance, and abnormal conscious or unconscious behaviour leading to accident or to deliberate self injury.

Readmission for mental disorder following an earlier admission for specific neurological diagnosis occurred in 233 cases, compared with an expected number of 77. Many of these readmissions may reflect deterioration of intellectual function or behaviour related to underlying organic brain disease. The high risk of readmission for mental disorders following a first admission for epilepsy (18 times that expected) is striking. It is also seen that there is a relatively high risk of readmission for mental disorder after head injury. This may be attributed to factors already discussed with, in addition, the possibility that admission may be due to changes in behaviour or mental function (for example, depression) occurring as a consequence of head injury or of the accident or violence which occasioned it. Subnormality caused by traumatic brain damage would also occur under this heading. It was found that admissions for other forms of trauma without mention of head injury were also followed by an excess of readmissions for mental disorder, namely 139 readmissions compared with an expected number of 49.1. It is likely that an examination of the secondary diagnoses listed on the medical record would reveal that in some of the cases readmitted for mental disorder after epilepsy or trauma, the mental disorder really coexisted in the former admission but was displaced to a secondary position because of the priority for treatment of the more urgent condition.

It is worth noting that a diagnosis of cerebral tumour was made on the second of a pair of admissions on 70 occasions when only 5.9

readmissions with this diagnostic label were expected. In 18 of these a change had been made only in the precise nomenclature of the tumour. Of the remaining 52 readmissions: seven had been preceded by an admission in which the primary diagnosis had been mental disorder, 12 had been preceded by an admission for stroke, nine by epilepsy, 13 by other specific neurological diagnoses and seven by neurological symptoms, three by cerebral trauma, and one by poisoning. A detailed study of the individual case notes would be necessary to determine whether diagnostic error was responsible for any of these associations. Some at least may be explained by mistakes in coding such as the displacement of a coexisting diagnosis of cerebral tumour in the earlier admission by an ill-defined condition such as 'hemiparesis'. In any event, the finding deserves further study.

ADMISSIONS TO HOSPITAL PRECEDING THE ADMISSION AT WHICH A NEOPLASM WAS DIAGNOSED

The extent to which patients admitted to hospital with malignant neoplasms may have been misdiagnosed or the presence of neoplasms overlooked during previous hospital admissions is a question of considerable clinical interest. In this study the method of calculation of the expected number of patients with any given diagnostic sequence makes it possible to estimate, for admissions in which the diagnosis of a malignant tumour has been made, the number of previous admissions due to any cause to be expected by chance in a group of persons of the same age and sex observed for the same length of time. The excess of admissions actually observed over those expected can then be calculated.

Table III shows the numbers of persons in whom selected neoplasms were diagnosed in hospital during the period of study together with the numbers who had admissions to hospital preceding that in which the specific diagnosis of the tumour was made. Admissions for other malignant tumours have been excluded from the preceding admissions because it has been assumed that these admissions were in fact mainly caused by the tumour specified at the later admission. Cancers of the colon and rectum have been grouped together. In the clinical context the preceding admissions of most interest are those for symptoms or diagnoses relating to the same system. Thus in the case of persons with primary cancer of the lung, Table III gives the number (and the proportion expressed as a percentage of all previous admissions in the study period) of earlier admissions of those persons ascribed to other diseases of the lung or pleura or to symptoms related to the respiratory system. For tumours of the alimentary tract Table III shows the number of previous admissions of the same patients for other gastroenterological conditions such as peptic ulcer or ulcerative colitis, or for alimentary symptoms such as upper or lower abdominal pain. Column 4 shows the number of previous admissions for such complaints as would have been expected in a group of patients of similar age who had not been admitted for a malignant tumour. Column 5 shows, by subtraction, the remainder. This surplus of unexpected admissions presumably relates both to admissions caused by the specific neoplasm but in which the diagnosis was not made, and to admissions associated with prodromata of the neoplasm. Of the tumours selected, benign and malignant tumours of the brain are seen to have the highest proportion

TABLE III
ADMISSIONS PRECEDING ADMISSION FOR SELECTED NEOPLASMS

Primary Tumour	Total Persons Admitted	Previous Admissions All Causes*	Previous Admissions from Causes* Related to Same System					
			Total		Expected		Unexpected	
			No.	%	No.	%	No.	%
Stomach	631	96	29	4.6	17.6	2.8	11.4	1.8
Rectum/colon	1247	304	82	6.6	32.8	2.6	49.2	3.9
Biliary system	113	28	8	7.1	2.7	2.4	5.3	4.7
Pancreas	246	53	16	6.5	6.3	2.6	9.7	3.9
Lung	1522	344	80	5.3	48.5	3.2	31.5	2.1
Breast	1204	192	14	1.2	10.4	0.9	3.6	0.3
Brain (benign and malignant)	371	170	51	13.7	5.0	1.3	46.0	12.4

*Excluding admissions for neoplasms

(12.4%) of unexpected previous admissions for conditions within the same system. Next in order of magnitude are tumours of the biliary system (4.7%), rectum and colon (4.0%), and of the pancreas (3.9%) but for these the proportions of unexpected previous admissions were much lower.

As far as cancer of the stomach is concerned 13 (2.1%) of the patients admitted had experienced a previous admission during the period which had been attributed to benign peptic ulcer.

The figures for cancer of the breast deserve particular comment. Although only 1.2% of cases had previously been in hospital for diseases of the female genital system, including all non-malignant conditions of the breast, as many as 15.9% of the women had been in hospital for other disorders during the study period, that is an average of 2.1 years. In some the diagnosis of breast cancer may have been made following careful routine examination of the breasts during inpatient treatment for an unrelated condition. In others a coexistent breast cancer may have been missed. A careful study of the clinical records and of the interval between the admissions is necessary before any conclusions can be reached.

READMISSIONS TO HOSPITAL AFTER DISCHARGE FOR TUBERCULOSIS

For the purpose of this analysis, all variants of tuberculosis disease contained in the *ICD* rubrics have been grouped together. Table IV shows the number of patients first admitted for tuberculosis and subsequently readmitted with other diagnoses.

A total of 1252 persons were admitted to hospital at least once for tuberculosis during the period covered by this study. Of these, 273 (21.8%) were readmitted with a second diagnosis different from that shown for the first admission, compared with an expected number of 166.7. Of the 273 cases readmitted, 24 were for another variant of tuberculosis. Of the remainder, the largest clinically distinct group consisted of 29 patients admitted for other forms of lower respiratory disease (excluding carcinoma), which was three times the expected number. Primary cancer of the lung was diagnosed in 18 readmissions, almost five times the expected number; in this case it is likely that the excess is due largely to the occurrence of diagnostic or coding errors at the time of the earlier admission—for example, the incorrect classification of a pleural effusion as of tuberculosis origin. However the possibility that patients with tuberculosis may have an increased risk of cancer of the lung cannot be entirely discounted. A hospital admission for tuberculosis is not followed by an excess of readmissions for all neoplasms other than cancer of the lung (there was a combined total of 20 cases which is close to the expected number of 17). Numbers for individual tumours are too small for meaningful interpretation but it is perhaps worth noting that there was an excess of readmissions after tuberculosis for tumours of the gastrointestinal tract (six observed, expected 3.2) and for cancer of the breast (four observed, expected 1.0).

TABLE IV
ADMISSIONS AFTER ADMISSION FOR TUBERCULOSIS*

Disease	Observed	Expected	Observed/ Expected	% of All Persons Admitted At Least Once For Tuberculosis
Other tuberculosis	24	0.4	60.0	1.9
Lower respiratory infection, other pulmonary, and pleural disease	29	9.0	3.2	2.3
Primary lung cancer	18	3.8	4.7	1.4
All other cancer	20	17.0	1.2	1.6
Gastrointestinal tract	6	3.2	1.9	0.5
Breast	4	1.0	4.0	0.3
Rheumatic heart disease	7	0.6	11.7	0.6
Arteriosclerotic, hypertensive, and other heart disease	16	13.0	1.2	1.3
Stroke	6	5.7	1.1	0.5
Mental disorder	21	5.0	4.2	1.7
Injuries	13	6.7	1.9	1.0
All other	109	101.3	1.1	8.7

*The total number of persons admitted once for tuberculosis was 1252. The average period at risk in terms of readmission was 2.1 years

For degenerative diseases (arteriosclerotic heart disease, hypertensive heart disease, cerebrovascular disease) the number of observed readmissions (16) after discharge for tuberculosis was similar to the expected number (13.0). However, for rheumatic heart disease, the observed number of seven is strikingly in excess of the expected number of 0.6. The converse situation was therefore examined (admission for rheumatic heart disease followed by readmission for tuberculosis), and an additional three cases of the combinations were found. This brings the total number of patients with separate admissions for both these diseases in either sequence to 10 as compared with an expected number of 1.2. The striking excess may be owing to the occurrence of tuberculosis and rheumatic heart disease in patients with common adverse social backgrounds. However, as the expression of both diseases is to some extent determined by atypical immune responses to bacterial infection further inquiry into the nature and extent of the association may be warranted.

Of the remaining conditions, readmission for mental disorder (21 cases) after tuberculosis is more than four times commoner than expected, while for injuries (13 cases) the excess is just less than twofold. All remaining diagnoses accounted for 109 admissions compared with an expected number of 101.3. Apart from the instances cited, there is therefore no evidence of any substantial excess of morbidity from other causes associated with tuberculosis.

DISCUSSION

Previous publications using Oxford Record Linkage Study material have measured the frequency of readmissions from home, and of transfer between hospitals, within a calendar year (Acheson and Barr, 1964) and within a two-year period (Fairbairn, 1968). It was shown that 22.7% of patients discharged were readmitted at least once within two years, and the frequency of readmission varied widely according to age, diagnosis, and other factors (Watts and Acheson, 1967; Baldwin, 1973). Using the same techniques, a study has also been made of the pattern of admission to hospital before and after certain common surgical operations (Fairbairn, 1972; Baldwin, 1973).

In a review of the research applications of cumulative personal medical records, Baldwin (1973) mentioned the ascertainment of spatial or temporal associations between pathological features in groups of individuals. It was with the possibility of demonstrating biological associations of epidemiological interest very much in mind that the linked

file was originally created (Acheson, Truelove, and Witts, 1961). In the material reported here, in which the average interval between the pairs of admissions is about two years, it is perhaps not surprising that no definite evidence of unknown associations of this sort has yet emerged. However, at least two hints which may be worthy of further study are mentioned in the text (the observed excess of admissions for breast cancer and for lung cancer after discharge for tuberculosis) and there may be others as yet unnoticed in the massive tabulation, of which the results reported here are but a small sample. One difficulty is that, without a prior hypothesis, an 'association' which in numerical terms is plain enough may be disregarded. For example, is the surplus of patients readmitted with mental illness after tuberculosis (21 observed; five expected) simply an indicator of common adverse social circumstances or could it imply a reaction to a drug? As is the case with many of the other questions raised, recourse to the medical records themselves ought to go a long way towards clarifying this point.

It is inevitable that an analysis of this sort will bring out associations which are due to social rather than biological factors. Associations due to age should generally have been excluded by the method used in calculating the expected numbers but no correction has been made for social class or occupation. An example of an association which might be due to common adverse social factors in this material is provided by the excess of admissions for rheumatic heart disease before and after admission for tuberculosis.

Berkson (1946) in a celebrated paper pointed to one of the pitfalls in drawing conclusions about the association of diseases from hospital inpatient material. He drew attention to the bias caused by a tendency for patients suffering from two or more diseases concurrently to seek hospital treatment more readily than patients with only one disease. This bias would make the combination of diseases appear spuriously to be commoner than would be expected from the prevalence of either disease separately. This bias probably applies to a lesser extent when the illnesses are consecutive as in this material rather than concurrent. However recognition of a second illness as a result of an examination undertaken at the time of admission for the first, will almost certainly lead in some instances to excesses of observed cases over the number expected. In his study of readmission after common surgical operations, Fairbairn (1972) attributed the excess of readmission for hydrocele after discharge for herniorrhaphy and for cystic

disease of the breast after hysterectomy and cholecystectomy to this effect.

Much of the interest of the associations noted in this paper falls within the clinical and administrative sphere. Attention has been drawn to the many occasions, particularly in the field of mental and neurological disease, where the differences between the diagnostic labels at the two discharges seem to represent two nosological versions of the same clinical phenomenon. To what extent these versions differ because of incorrect diagnostic coding or because of the availability of new information at the second admission would require a detailed study of the case notes. It is hoped that a more sophisticated analysis taking account of differential mortality, emigration, and other factors in the computation of the expected numbers together with subsidiary diagnoses will be published in due course.

A second finding of considerable clinical interest has been the high proportion of patients in whom there has been a preceding admission to hospital under some other diagnostic labels but involving the same physiological system before a first definitive diagnosis of a malignant lesion. In a proportion of these admissions, the malignant lesion was probably present at the time of the earlier admission. This material gives an opportunity for a systematic appraisal of the hospital precursors of malignant disease. This is an important facet of the wider subject of the natural history of disease which may be studied conveniently by linked records.

We are greatly indebted to Dr Jack Howlett, Director of the Atlas Computer Laboratory at Chilton Down and to his staff for the generous facilities put at our disposal and for their help throughout. We also thank Mr Leicester Gill, Computer Scientist in the Unit of Clinical Epidemiology, for his important role in the preparation of the matrices.

Requests for reprints: Professor E. D. Acheson, Southampton General Hospital, Community Medicine and Primary Medical Care, South Academic Block, Tremona Road, Southampton SO9 4XY.

REFERENCES

- ACHESON, E. D. and BARR, A. (1964). Multiple spells of inpatient treatment in a calendar year. *Brit. J. prev. soc. Med.*, **19**, 182.
- , TRUELOVE, S. C., and WITTS, L. J. (1961). National epidemiology. *Brit. med. J.*, **7**, 668.
- BALDWIN, J. A. (1973). Linked record medical information system. *Proc. roy. Soc. B*, **184**, 403.
- BERKSON, J. (1946). Limitations of the application of fourfold table analysis by hospital data. *Biomet. Bull.*, **2**, 47.
- DALGARD, O. S. (1973). *Migration and Functional Psychoses in Oslo, Norway*. University Press, Oslo.
- FAIRBAIRN, A. S. (1968). In *Record Linkage in Medicine: Proceedings of the International Symposium, Oxford* edited by E. D. Acheson. Livingstone, Edinburgh.
- (1972). Hospital admissions following common surgical operation. *Proc. roy. Soc. Med.*, **65**, 239.
- FEDRICK, J. and BALDWIN, J. A. (1976). Methodological considerations in analysing a file of linked diagnostic data. (In preparation).
- WATTS, S. P. and ACHESON, E. D. (1967). Computer methods for deriving hospital inpatient morbidity statistics based on the person as a unit. *Brit. med. J.*, **4**, 476.

APPENDIX

METHOD OF CALCULATING OBSERVED AND EXPECTED NUMBERS OF DIAGNOSTIC COMBINATIONS

In any calendar period, a patient may incur more than one hospital discharge. The greatest number of discharges in five years in one patient was 34. In any sequence of discharges the same diagnosis may be repeated. In order that one patient should not contribute more than one to the observed number for a specified sequence of two diagnoses, the following procedure was adopted. Suppose that a sequence of diagnoses AAABCBAD was encountered in one patient and that A, B, C, and D represent different diagnoses. Repetitions of the same diagnosis were first eliminated, reducing the string to ABCD. This reduced string then contributed one to the count for each of the six diagnostic sequences AB, AC, AD, BC, BD, and CD. All hospital admissions have been included in this procedure and no distinction has been drawn between discharges home and interhospital transfers.

Each expected number was calculated by compound probability. The expectation was calculated separately for each age-sex-area group and the separate expectations added together to provide a total expectation, standardized for age, sex, and area for the diagnostic sequence A, B. Thus for the Oxford area, with five years' observation:

- Let p_{AB} = probability of one individual in the population incurring an admission with diagnosis A followed by an admission with diagnosis B. Similarly,
- p_{BA} = probability of diagnosis B followed by diagnosis A.

p_A, p_B	=	separate probabilities of an individual incurring diagnoses A, B in five years.
P	=	average population of the area during the five years
n_A, n_B	=	total number of hospital discharges with diagnosis A, B in five years.
Then p_A	=	n_A/P and $p_B = n_B/P$
and $p_{AB} + p_{BA}$	=	probability of incurring both diagnosis A and diagnosis B irrespective of order.
	=	$p_A \times p_B$ by compound probability
Hence p_{AB}	=	$p_{BA} = \frac{1}{2} \times p_A \times p_B$
Let E_{AB}	=	the expected number of individuals in the population incurring an admission with diagnosis A followed by an admission with diagnosis B.
The E_{AB}	=	$P \times p_{AB}$
	=	$\frac{1}{2} \times P \times p_A \times p_B$
	=	$\frac{1}{2} \times P \times \frac{n_A}{P} \times \frac{n_B}{P}$
	=	$\frac{\frac{1}{2} \times n_A \times n_B}{P}$

The calculation for the Reading area, with only two years' observation, is exactly analogous to that for Oxford.

Four tables, one for each sex-area group, of the numbers of discharges by diagnosis and age were constructed by a separate run of the computer and held on magnetic tape. These tables were transferred to the computer store before the computation and so could be accessed to provide the values of n_A and n_B . Similarly, tables of population by age, sex, and area were constructed from data published by the Office of Population Censuses and Surveys and stored in the computer. The required value of P could be obtained by accessing these tables. Finally, all $2 \times 2 \times 20 = 80$ values of E_{AB} were summed to yield an expected number, standardized for age, sex, and area for every diagnostic sequence A, B.

APPROXIMATIONS AND ASSUMPTIONS

In computing the expected number of patients in whom diagnosis A is followed by diagnosis B, the assumption is made that in one age-sex-area group the probabilities of both diagnosis A and diagnosis B are uniform throughout the population and throughout the calendar period of observation. This assumption is untrue under the following conditions:

1. If diagnosis A carries a risk of mortality. This objection is of particular importance when diagnosis A is for malignant disease, or for cardiovascular accidents, when the comparison of observed and expected numbers is quite invalid.
2. If some of the patients discharged with diagnosis A emigrate from the area before the end of the calendar period of observation, they would have a reduced chance of readmission and discharge with diagnosis B within the study area. Thus emigration would tend in general to reduce the observed numbers of patients with any disease combination. The expected numbers are calculated from the numbers of discharges for the two diseases and the numbers of persons in the population, and these are for the most part relatively unchanged from one year to the next. Where this is so it may be assumed that a patient who emigrates is replaced by one (or in this area more than one) immigrant of the same age and sex with the same risk of disease. In these cases therefore the expected numbers are not attenuated by the effects of migration as are the observed numbers. Hence it may be argued that if a striking excess of observed cases over expectation is found, such excess would have been even more striking but for the effects of migration. It is possible also that, because of this reduction in observed numbers, some findings of importance might be missed through failing to satisfy the criteria for inclusion in the shorter list. There is a dearth of information on the extent of migration into and out of the study area, but an inquiry carried out by the Oxford City and County Executive Council into removals and re-registrations with general practitioners indicated that migration varies with age. Among the older age groups, which are most affected by chronic diseases and would be most likely to have high readmission rates, the loss from the study caused by

emigration is probably about 2% a year, giving a total loss during the five-year period of about 10%. Emigration rates were very much higher in young adults, and losses of the order of 50% during the five years are not unlikely. In these groups particularly, the population has been increasing rapidly and numbers of hospital discharges have increased accordingly. A further possible complication is that patients with certain diseases may be subject to unusual migration rates, such as has been found elsewhere for some mental disorders (Dalgard, 1973).

3. If there is a secular trend in the frequency of disease A or disease B during the period of observation, both the observed and expected numbers of cases would be affected. Data on secular trends are available from the study records, but correction for this variable would be complex and was not attempted for this preliminary analysis.
4. The method assumes that illnesses occur evenly throughout the year and do not cluster together in time. If there is a seasonal variation in the frequency of disease A or disease B and the point of occurrence or the interval between the diagnoses is not constant, the probability of the combination occurring will vary throughout the calendar year. An example might be cancer of the lung followed by bronchopneumonia. The method will not detect associations of this kind.
5. The expected numbers are calculated on the assumption that within one age-sex-area group all persons have the same probability of discharge from hospital with diagnosis A or B. However, discharges with a particular diagnosis are not randomly distributed within the group but are clustered in certain patients who have repeated admissions and discharges with the same diagnosis. The procedure adopted in the calculation of observed numbers disregards any discharges with the same diagnosis as in a previous discharge of the same patient. In contrast, the expectations are calculated from the overall frequency of diagnosis A or diagnosis B including diagnoses which are repeated in different discharges of the same patient. As with the effects of migration, this discrepancy will tend to inflate the expected number of

cases and so narrow any gap between observation and expectation, thus obscuring some positive findings.

The method as applied to the data for this study also involves certain approximations as follows:

1. The method assumes that all cases having diagnosis A or diagnosis B are taken into account in deriving observed and expected numbers. In this preliminary analysis only principal, primary, or first mentioned diagnoses were included so that much diagnostic information was omitted. This could have caused low values for either observed or expected numbers, or both, and should be borne in mind in interpreting the results.
2. Since diagnoses are ascertained by the study only on discharge, records of patients readmitted and remaining in hospital throughout the balance of the follow-up period would not be available to it. As with patients who emigrate, the observed numbers with diagnosis B would be reduced. A further possible difficulty arises where there is likelihood of long-term hospitalization in that the point of discharge may be unrelated to the point of incidence of disorder so that the discharge rate is not an appropriate measure of disease incidence. This situation is likely to be important only in limited groups of diseases, such as some mental disorders.
3. The calculation is based on five-year age groups so that no allowance is made for the movement of patients from one age group to the next during the follow-up period. This 'cohort effect' can be controlled by using year of birth groups but this was not attempted in this initial analysis.

The effect of most of the assumptions and approximations made for this analysis is to reduce the ratio of observed to expected numbers. Marked increases in observed over expected numbers are therefore reported with some confidence. Conversely, it must be emphasized that some associations between diagnoses may have been missed, while smaller than expected numbers should not be accepted as reliable. A method of analysis has been developed subsequent to this study to take account of the assumptions and approximations so far as the data permit and results will be reported later (Fedrick and Baldwin, 1976).