AUTOMATIC LINKAGE OF MEDICAL AND VITAL REGISTRATION RECORDS

BY

ALWYN SMITH

From the Scottish Home and Health Department, the Office of the Registrar General for Scotland, and the Department of Public Health and Social Medicine, University of Edinburgh

The desirability of being able to link two or more records of vital or medical events occurring to the same individual is largely self-evident. In clinical practice it is often informative to consult past records of hospital or other treatment and it is often useful to be able to locate subsequent records in order to follow up patients treated at some time in the past. Research applications of such record linkage are numerous. They arise, for example, when the research worker wishes to examine associations between past and present morbidity experience and where the possibility of biased recollection might invalidate information obtained by interview. For example, the question whether *x*-irradiation of the pregnant mother raises the risk of leukaemia in the offspring of the pregnancy remained unresolved for several years because of doubts concerning the validity of data on radiography during pregnancy based on interviews with the mothers of affected and unaffected children. If data may be conveniently retrieved from contemporary records such doubts may be eliminated.

It may eventually be possible to maintain a central medical file for each individual in the community. This would be initiated at birth, or possibly before, and kept up to date by the systematic addition of contemporarily-made records of all relevant events in his life. If this does become possible, we shall need to be able to identify each individual uniquely and automatically. Meanwhile a more modest approach involving *ad hoc* record matching will also require a suitable identification device capable of being recognized by conventional statistical machinery or by electronic computer.

The requirements of such an identification device are that it should uniquely, or virtually uniquely, specify the related individual, that its form should be consistent and readily recognizable, that it should

be readily obtainable from the individual, and that it should be easily recorded and processed on punched cards or on magnetic or paper tape.

In a paper given to the Society for Social Medicine in 1962, Dr. E. D. Acheson reviewed available systems with reference to the compilation of a register being established in the Oxford Region where a computer will be available. What follows owes much to Acheson's paper but gives consideration to the special circumstances of Scottish national statistics where computer facilities are unlikely to be available on such a lavish scale as in Oxford.

Three main means of identification are worthy of consideration:

- (1) Conventional identification, using the name, address, date and place of birth, and other similar information obtainable from the individual.
- (2) Numerical identification using an existing number unique to the individual (*e.g.* the N.H.S. number).
- (3) Numerical identification by a number created for the purpose and allocated to each individual.

The advantages of conventional identification include its relatively easy recording from information supplied by the individual, and its general reliability. Its disadvantages are mainly that it is cumbersome to use in automatic data processing and that it may not uniquely specify the individual. The advantage of using existing numerical identifications is mainly that they exist. Their disadvantages are that most of them are far from universal, they are usually inconsistent in form, and are largely unknown to the individuals they identify. Clearly, many of these disadvantages could be avoided in a number expressly designed for the purpose. It would however, retain the disadvantage that few of the

identified individuals could be relied upon to remember their number accurately so as to supply it at the time of hospital admission or at the recording of some other relevant event. It becomes clear that an ideal identification would be one designed for the purpose and obtainable from the individual. But before considering whether this might be possible, it will perhaps be useful to examine possible existing systems.

In Great Britain there are two identification systems already in existence affording a wide coverage of the population. These are the National Health Service number, and the birth registration entry number. In respect of persons born since September, 1939, the two identifications are substantially the same. The National Health Service number of an individual is in most cases the number he either had or would have had on a wartime Identity Card or Ration Book. In September, 1939, the population was enumerated as at a census and each individual acquired a number which consisted of a code for the district he was in at enumeration, a code for the household or other establishment within the district, and a number identifying him within the household. Thus the writer's own number was OOAB 131/3; the alphabetical digits identify the district, the numbers to the left of the oblique stroke the household, and the number to the right of it the individual as the third person enumerated within the household. The alphabetical identification of districts followed a United Kingdom system, all Scottish district codes beginning with "S". The household identifications were arranged so that large institutions were mostly given single digit identifications (e.g. Edinburgh Royal Infirmary was "1"). This obviated unmanageably large numbers for people in such institutions at the registration date. Unfortunately, variation in the number of digits makes the oblique stroke a vital part of the number. A person having the identification SXYZ 1/23 would be different from a person whose number was SXYZ 12/3. Thus any system for transferring the number on to punched cards or tape must specify the location of the oblique stroke.

Persons who were in H.M. Forces on the registration date or who joined the forces and were demobilized before the end of food rationing, acquired a different number. This consisted usually of a three digit alphabetical district identification followed by seven numerical digits. Thus many people have had two numbers of different form.

In the case of persons born in the United Kingdom since September, 1939, a different system has been in use and the systems used in England and in Scotland differ from one another. In both countries the

number identifies the person's birth registration entry, but the identification differs between England and Scotland and to some extent also within Scotland. In England the birth register is a book carrying a four digit alphabetical identification which is unique to the book. A new book is started when the current one is full, irrespective of the date, and each book contains less than 1,000 names. Thus all English persons born since September, 1939, have a N.H.S. number having the general form ABCD 123. In Scotland, however, one book is used for each registration district for each year, so that the books vary considerably in size. They are identified by the code number of the registration district followed by the year to which they refer. Moreover, registration districts within large Local Authority areas have numbers consisting of a three digit code identifying the Local Authority followed by either a numerical or alphabetical code for the individual district. This district code may have one digit or two. N.H.S. numbers are thus made up of a variable form of district identification, followed by the year of birth (sometimes only the last two digits), followed by the entry number in the book of that district for that year, and always preceded by the letter "S".

Thus a Scottish N.H.S. number may take forms as various as S 644.19/1962/9987 or S 322.a/60/1 or SAAA/1/1. The problems of arranging such numbers intelligibly on a punched card are formidable.

But apart from the inconsistency of form, there is the problem that an individual's number is not easily obtainable. Few people know their numbers, and few carry their medical cards. The number should be recorded on the general practitioner's record, at the Local Executive Council, and at the Central Register of the National Health Service, but in practice Acheson failed to obtain it for a substantial proportion of all patients.

Many of the objections to the N.H.S. number for use as an identification could be countered if the number were reconstructed on a consistent basis, and if it were more widely used (and therefore more accessible). Its advantages include its existence, its uniqueness, and its meaning in so far as people born since September, 1939, are concerned.

The problem of creating a new identification is similar to that of reconstructing the N.H.S. number. As Acheson has pointed out, the most important point is that to obtain an individual's number subsequently it would usually be necessary to consult a file using conventional identification information obtained from the person or his relatives.

This makes it natural to consider the possibility of using this conventional information directly. Acheson favours this approach and it is undoubtedly

possible if a powerful electronic computer is available. Even so, there are many difficulties in ensuring a sensibly economic use of computer time. Hogben, Johnstone, and Cross (1948) have already proposed a system whereby easily available information is coded into numerical form to facilitate its use in conventional data processing machines. A code identification is derived by translating the individual's given name and surname into a four digit sequence which distributes names in proportion to their frequency of occurrence; this name code is followed by the date of birth written in numerical form (1st January, 1963 becomes 01 01 63; 15th November, 1962 would become 15 11 62). This code has the merit that one can construct it consistently from the patient's name and date of birth but one cannot proceed in the reverse direction and name a person from his code number; thus the information remains confidential. Unfortunately patients do not give their dates of birth with great consistency. About 10 per cent. of people admitted twice to hospital give different birth dates at the two admissions (Acheson, 1962; Innes and Sharp, 1962). Differences are mainly in the final digit of the year but also occur at other places. Inconsistency is also encountered in the recording of names. Where an individual has two given names the order in which he records them may not be consistent. In particular, he may not usually be called by his first name so that on one occasion he may give the name by which he is called, and on another occasion his actual first name. Many given names have diminutive or popular forms differing radically from the orthodox form. Thus the code identification derived from a patient may differ on different occasions and simple machine matching of the records may fail.

The problem then is to devise an identification code which depends upon information obtainable from individuals, which may be readily coded in a consistent form and which is virtually unique to the individual. We have then to process the information so that allowance may be made for an element of unreliability in the recorded data. The latter requirement involves the making of intelligent decisions concerning incompletely matched identifications. Two commonly employed devices are able to make such decisions; human clerks and electronic computers; but both are in practice limited in their ability to search for matching records among large volumes of data. Computers are limited by the size of their memories which become enormously expensive as they become larger. Human clerks are limited by the speed at which they operate as well as by the size of their memories which are relatively small so far as concerns the kind of data involved in

medical records. Conventional punched-card machines are relatively cheap, and can sort large quantities of data, but have very limited powers of intelligent decision.

The limitations of conventional machinery would be crucial in dealing with name/date codes such as that of Hogben since an error in one digit would usually suffice to prevent a successful automatic linkage. But, if we could use conventional machines to sort the data within a system having a high degree of reliability, the residual matching could be enormously facilitated since the searching would be greatly reduced. In practice, computer methods for data linkage usually rely on preliminary sorting of this kind so that modestly powerful computers may carry out the residual matching. It seems probable that an analogous method would permit the clerical matching of data, at least on a modest scale.

The problem now becomes one of coding the reliably recorded data for residual matching. Clearly the specificity of the reliable part of the code should be as high as possible so as to reduce the search and decision-making with respect to the rest.

In the Hogben system it is probable that surname coding is more reliable than given name coding but that it is insufficiently specific for our purpose. The Hogben system converts surnames into a two-digit code, so arranged that each of the 99 codes occurs with approximately equal frequency. His own code was simply constructed by dividing the Birmingham telephone directory into 99 approximately equally-spaced sections and allocating a code to each of the alphabetical groups of names so distinguished. Thus names beginning Aa to Ak are coded 00; Al is coded 01; Am to Ar is coded 02 . . . Si to Sm is coded 79; Smith is coded 80 etc. For Scottish names different codes are required.

A defect of the Hogben system derives from the unreliability of given names and the changes in fashion which alter their frequency distribution. The present proposal is therefore to use a paired parental-name code in which the two-digit code deriving from the individual's surname at birth is followed by the two-digit code derived from his mother's maiden surname. The use of paired parental names for identification purposes is not new, the method having already been employed by Newcombe and his colleagues in British Columbia for linking marriage and birth records (Newcombe and Rhynas, 1962). What is new is the proposal to use paired parental surname codes based on a simple two-digit code such as that proposed by Hogben and others (1948), and to extend their use to various linkages required in national health, medical, and vital registration records. Newcombe and Rhynas employ a more

complex name coding in which both parental surnames are converted into the Russell Soundex Code. This code is more specific than Hogben's, having four digits for each surname, and it has the advantage of being somewhat less sensitive to possible inconsistencies in spelling. However, our experience is that in Scotland surname recording is reasonably consistent and the Soundex code would offer no important advantages to offset the disadvantage of greatly increased coding and sorting time. There is also reason to believe that the four-digit paired-surname code we propose to use will be adequately specific for our purpose. There is no evidence that assortative mating for surnames is important in Scotland as it may be in North America with its large immigrant groups. Thus the four-digit code should distribute people into 10,000 more or less equal groups. The population of Scotland is about five million, so that a group of individuals having a common paired-name code might contain an average of 500 persons. In most practical applications the number will be very much smaller. For example, there are approximately 100,000 births per year and thus only about ten births per year will fall into any one paired-name category.

The paired parental-surname code will consist of the two-digit code of the father's surname followed by the two-digit code of the mother's maiden surname. Where either is not available the code 00 will be used. This permits a meaningful code to be attached to illegitimately-born individuals and will also permit identification of cases where an alternative search may need to be made for a matching record.

A suitable coding of Scottish surnames has been prepared from data obtained from vital registration (Table, opposite).

This has been constructed so that, as far as practicable, each two-digit code contains an equal number of individuals and only the first two characters of the surname are needed for coding. Several exceptions need to be made to this rule; in some cases the first three characters have been used, and several common names require codes to themselves. The Scottish name prefix "Mac", "Mc", or "M" has been counted as a single character.

A question needing to be answered concerns the availability of the mother's maiden surname. This is, of course, rarely recorded on hospital documents but there seems no good reason why it should not be. It seems likely that it could more easily be obtained than the N.H.S. number. An attempt is currently being made in the Aberdeen Mental Hospital Group and in Dundee mental hospitals to record this information in respect of all referred patients. Other

similar studies will be initiated elsewhere. The information is requested at registration of all vital events in Scotland and is virtually invariably obtained in respect of registration of birth, stillbirth, and marriage (for both parties). Death registration differs in that the informant is sometimes unrelated to the deceased and the deceased is sometimes very old. In addition, for married women, two names are recorded other than that of the deceased person. Even so, the maiden surname of the deceased's mother is obtained in 98 per cent. of registrations, and the maiden surname of married women in 99 per cent.

The data needed for residual matching will vary according to context and to what is available. When residual matching is to be performed clerically with punched cards, selected data will be interpreted (*i.e.* printed by a machine which reads selected columns and prints their contents along the top of the card).

The procedures may best be illustrated by examples.

(1) *Matching of Birth and Stillbirth Records to the Related Marriage*

At marriage registration, the names of both parties are recorded, with the date and place of marriage. At legitimate birth and stillbirth registration the name of the father as well as the maiden surname of the mother is recorded, with the date and place of marriage of the parents. The marriages in one year are sorted into four-digit paired-name code order and births in subsequent years are similarly sorted; matching is then achieved clerically by searching corresponding batches. This can be achieved even where the date or place of marriage is wrongly recorded if additional data are consulted as the basis for a decision. Such additional data might well be the parental ages, the given names, the registration district, and the occupation of the father. None of these items alone will permit an unchallengeable match, but together they will permit a highly probable one. A trial along these lines is proceeding in respect of registrations for Aberdeen City and County, and the possibility of instituting a national system is being investigated.

(2) *Matching of Records Relating to Successive Hospital Admissions of the Same Individuals*

The admission data would be kept on punched cards and sorted into hospital code order followed by paired-name code order. The recorded data at subsequent admissions should then include the last hospital to which the individual was admitted, and these new cards, sorted similarly, could be readily matched with those relating to the previous admission. Each new admission card would be placed

behind the discharge data cards relating to the previous admission of that person. Annual tabulation and production of summary punched cards would permit analysis of hospital data on the basis of persons rather than of admissions or discharges. Scottish mental health statistics are currently being compiled in this way, except that at present given name codes are being used instead of mother's surname. For larger scale data the paired parental-name code would facilitate final matching, which may take into account other recorded data such as diagnosis, age, occupation, etc.

The procedure described above is capable of many applications and may conveniently be adapted for residual matching with either clerical workers or electronic computers. The decision processes are essentially similar in either case.

The identification code proposed by Hogben and others (1948) was constructed by adding to the four-digit name sub-code the six-digit numerical

form of the date of birth. The given name code incorporated a sex distinction in that male given names were allocated codes 00 to 49 and female given names 50 to 99. The present system lends itself to a similar application, except that an additional digit would be needed for sex differentiation. But it is doubtful whether a unique code number is either necessary or desirable at the present time when machine handling can be carried out with far more sophistication than was possible when Hogben's system was proposed. The present system permits the employment of whatever residual information seems most appropriate. Date of birth will often be used but where fraternities need to be assembled the date of the parents' marriage would obviously be a more useful item.

The applications of a system such as this to vital registration data are of course potentially far-reaching. At marriage, the paired parental-name codes of each partner and their dates of birth as

TABLE
TWO-DIGIT SURNAME CODE FOR SCOTLAND
(Alphabetic Code Listing)

AA-AK	01	FM-FO	28	MAC/McA, B	50	RA-RD	76
AL-AM	02	FP-FZ	29	CA-CL	51	RE	77
AN-AN	03			CM-CZ	52	RF-ROB (Ex 79)	78
				D	53	ROBERTSON	79
BAA-BAQ	04	GA-GH	30	E, F	54	ROC-RZ	80
BAR-BD	05	GI	31	G	55		
BE	06	GJ-GQ	32	H-J	56	SA-SC	81
BF-BN	07	GRA	33	KA-KE	57	SD-SH	82
BO-BQ	08	GRB-GZ	34	KF-KZ	58	SI-SL	83
BROWN	09			LA	59	SMITH	84
BR (Ex 09)	10	HAA-HAQ	35	LB-LZ	60	SM-SS (Ex 84)	85
BS-BZ	11	HAE-HAZ	36	M, N	61	STEWART	86
		HB-HH	37	O-Z	63	ST (Ex 86)	87
CAA-CAM (Ex 13)	12	HI-HOL	38			SU-SZ	88
CAMPBELL	13	HOM-HT	39	MAA-MAQ	63		
CAB-CG	14	HU-HZ	40	MAR-MAZ	64	TA-TG	89
CH-CN	15			MB-MH	65	TH	90
COA-CON	16			MI-MN	66	TI-TZ	91
COO-COZ	17	I	98	MOA-MOO	67		
CP-CR	18			MOP-MT	68		
CS-CZ	19	JA-JN	41	MUA-MUN	69	U, V	99
		JO-JZ	42	MUO-MZ	70		
DA-DD	20					WAA-WAQ	92
DE-DN	21			N	71	WAR-WAZ	93
DOA-DON	22	KA-KEL	43			WB-WH	94
DOO-DT	23	KEM-KEZ	44	O	72	WILSON	95
DU-SZ	24	KF-KZ	45			WI (Ex 95)	96
				PA-PD	73	WJ-WZ	97
E	25	LA-LD	46	PE-PN	74		
		LE-LH	47	PO-PZ	75	X-Z	99
FA-FH	26	LI-LN	48				
FI-FL	27	LO-LZ	49	Q	98	NONE	00

well as the date of marriage would be recorded. At birth and stillbirth registration, the paired parental-name codes of the newborn and his date of birth, as well as the date of his parents' marriage would be recorded. At death registration, the paired-name codes and the date of birth would be recorded. When several generations had been recorded in this way, automatic pedigree construction by electronic computer would be a relatively simple matter. Human genetical research has hitherto been compelled to adopt a methodology wholly different from that developed in animal genetics because of the different circumstances in which human records have been obtainable; there can be little doubt of the potential value of information stored so as to permit the automatic searching of records relating to several generations. Moreover, it is not even necessary to await the compilation of records for several generations hence. There is no reason apart from the expense why Scottish registration data back to 1854 should not be abstracted and recorded in this way. An immediate application is being instituted in respect of a study of familial recurrence of stillbirth using data from 1939 to 1962.

But retrospective data retrieval is not the primary application the writer wishes to stress. The use of this system seems to have many advantages for the linkage of current data from vital registration and from health and morbidity statistics. One obvious advantage is that the information would remain confidential, since the identification code permits the machine-matching of related records without permitting the identification (in the personal sense) of the individual. If the codes only were recorded on the statistical documents and the full identification as well as the codes were recorded on the related clinical record, the statistical record would remain anonymous in the conventional sense while the related record could still be located by the doctor in charge of the patient. Even if (as seems best in practice) the coding is performed centrally, statistical returns need contain only the first few letters of each parental surname and would thus remain effectively anonymous. The method seems to offer

the possibility of simple semi-automatic linkage with existing statistical machines as well as convenient adaptation to electronic computer processing when such machines become available. Extensive trials and development are proposed and some are already in progress.

SUMMARY

There is need for a convenient numerical identification device to permit the automatic or semi-automatic linkage of related medical and other records.

Existing identification numbers are generally unsuitable; the N.H.S. number is particularly difficult to handle.

A proposal is made to use a four-digit preliminary sorting device consisting of paired two-digit sub-codes deriving from the individual's parents' surnames. Residual matching within the four-digit categories so determined may be carried out either clerically or by electronic computer, using whatever information is best suited to the purpose.

The possibility of automatic pedigree construction from vital registration records is envisaged.

I wish to acknowledge my indebtedness to the staffs of the Registrar General for Scotland and the Scottish Home and Health Department for the preparation of material, and to many colleagues for their valuable criticism.

REFERENCES

- Acheson, E. D. (1962). Paper given to the Society for Social Medicine, September, 1962.
- Hogben, L., Johnstone, M. M., and Cross, K. W. (1948). *Brit. med. J.*, 1, 632.
- Innes, G., and Sharp, G. A. (1962). Personal communication.
- Newcombe, H. B., and Rhynas, P. O. W. (1962). In "The Use of Vital and Health Statistics for Genetic and Radiation Studies", Proceedings of the Seminar sponsored by the United Nations and the World Health Organization Geneva, 1960, p.135. U.N.O., New York.