J Chron Dis Vol. 34. pp. 291 to 297. 1981
Printed in Great Britain. All rights reserved

0021-9681/81/060291-07$02.00/0

COHORT FOLLOW-UP USING COMPUTER LINKAGE WITH ROUTINELY COLLECTED DATA

D. J. HOLE*, J. A. CLARKE†, V. M. HAWTHORNE‡ and R. M. MURDOCH§

*West of Scotland Cancer Surveillance Unit, Ruchill Hospital, Glasgow, †Information Services Division, Scottish Health Service, Common Services Agency, Edinburgh, Scotland, ‡Department of Epidemiology, University of Michigan, School of Public Health, Ann Arbor, Michigan 48109, U.S.A. and §Department of Community Medicine, University of Glasgow, Ruchill Hospital, Glasgow, Scotland

(*Received in revised form* 13 *June* 1980)

Abstract—The validation of a computerised record linkage system for matching members of a defined cohort with routinely collected national data sources is reported for the first time. The two national sources relate to mortality and inpatient data and provide contrasting characteristics in their method of collection. The linkage system produces a group of possible matches based on identifying information restricted to surname, initials, sex and date of birth which, with a clerically assisted scrutiny, gives levels of sensitivity of 66% for the mortality and 81% for the inpatient data. Specificity can be increased to 100% if conventional follow-up methods are used for the limited set of matches classified as probable by the clerical scrutiny.

INTRODUCTION

THE COHORT study is a major methodological technique in epidemiological investigation. Long term follow-up of subjects to identify appropriate end points can provide information to test hypotheses about groups within the cohort with a particular characteristic or exposure to an occupational, environmental or other potential hazard.

Present methods of following individuals range from personal contact to tagging in appropriate data sources such as nationally collected data files. The problems involved include declining response rates at re-examination, the cost and time needed to conduct re-examination on large populations and the accuracy and completeness of manual tagging systems.

With the advent of national computerised medical records, it is possible to use automated linkage procedures to improve the completeness of follow-up and to economise in resources. However, the use of a computerised system for matching records does not include all the subjective judgements which could be made when comparing non-identical identifying information. At the present state of development of automated matching, the best method of allowing for variability of identifying data seems to be clerical scrutiny of the limited number of linkages provided by the computer.

The effectiveness of any linkage system can only be ascertained by examining its performance with different sets of data. This paper evaluates the effectiveness of one such system which includes a clerical check, in matching data directly collected from a cohort with two sets of nationally collected data.

MATERIAL AND METHODS

The study cohort was a sample of 3062 residents from the Scottish town of Renfrew, aged between 45 and 64 yr and examined in 1972 for a large scale prospective study of 'asymptomatic' cardio-respiratory disease by Hawthorne *et al.* [1]. Each member of the cohort gave written permission for his or her medical record to be examined by the study team for research purposes.

The two sets of nationally collected data were the Scottish mortality and inpatient files. The mortality data are derived from death certificates held by the General Register Office (Scotland) and for this study related to all deaths occurring in Scotland between 1972 and 1974 inclusive. These are coded centrally by the G.R.O. directly from the death certificates. The hospital inpatient data, collected on the statistical abstract form SMR1, contains all discharges from Scottish hospitals (excluding mental hospitals and maternity units) and information for the year 1973 was used. The SMR1 document is completed by medical records clerks at each hospital taking information from the case record. These data are processed by the Scottish Office Computer Service for the Information Services Division of the Common Services Agency. Heasman [2] has described the characteristics and content of this material.

The linkage method employed was derived from the work of Newcombe and his colleagues [3] in Canada, Acheson at Oxford [4] and is summarised in a paper by Phillips [5]. It used as identifying information the surname, first two initials, sex, day, month and year of birth.

Comparison is made on the similarity of the Russell-Soundex phonetic representation of the surname and the equivalence of the other items. A probability weighting is assigned according to the correspondence of the compared items and their value in establishing a match. A detailed description of the weighting procedure is given in a paper by Sunter [6]. The probability weightings are converted into numerical match weights for all possible pairs of records. A listing can then be produced of all pairings above any given threshold match weight. In this study a match weighting of 140 was selected because pilot studies suggested that this level would produce all probable matches.

The terminology used in this study is defined below (Fig. 1). A pair of records whose weight exceeds the given threshold is called a ‘match’. In some matches the identifying data are the same and these are described as ‘identical’ matches. Both ‘identical’ and ‘non-identical’ matches are described as ‘correct’ if the pair of records refer to the same person.

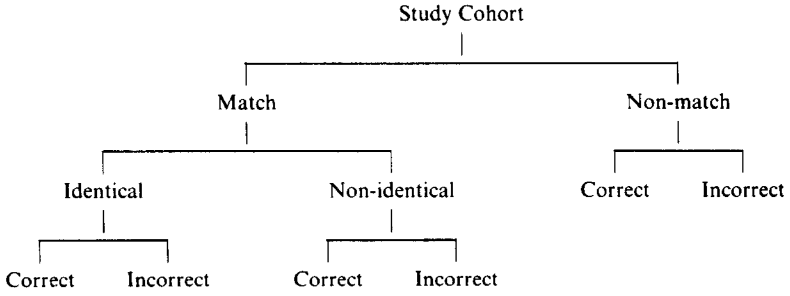

Fig. 1.

For the ‘non-matches a record is defined as ‘incorrect’ if it is found from other sources that a death or inpatient stay has occurred for that person.

The correctness and completeness of the mortality matches were established by comparison with deaths reported as a result of manual tagging of the study cohort in the National Health Service Central Register (NHSCR).

The correctness of hospital inpatient matches was assessed by one observer checking a stratified random sample of the matched cohort records with the hospital case records indicated on the matching SMR1.

The completeness of the hospital inpatient matches was assessed by examining a random sample of the non-matches using their general practitioner records to determine whether any hospital discharge had occurred. A subject with such a recorded discharge was thus described as an incorrect non-match.

The clerical checking procedure was carried out in the following manner. Three observers independently judged the correctness of the computer matches and allocated each match to a category—positive, equivocal, improbable or negative. To discover whether a combined judgement was superior to the individual ones, arbitrary weights of 5, 3, 2, 0 were assigned to the respective decision categories. A combined score between 0 and 15 was then obtained for each match by summing the scores of the three observers. Arbitrary cut points were selected to divide these combined scores into the four decision categories analogous to those used by individual observers. The decisions reached by the three observers, taken both individually and collectively, were then compared with the results obtained using the manual tagging (mortality) and the check with hospital case records (inpatient discharges).

The criteria on which a match was accepted was varied to allow an examination of the way in which the levels of sensitivity and specificity of the linkage system changed.

RESULTS

(a) *Computer Linkage*

Linkage of the cohort with Scottish mortality data for 1972–74 (containing 194,302 entries) generated 319 matches above the 140 threshold. Table 1 shows the breakdown of these matches and their accuracy.

TABLE 1. ACCURACY OF COMPUTER LINKAGE FOR MATCHES OBTAINED WITH 1974 MORTALITY FILE

	No. in	No. traced and verified as			
Type of Match	Cohort	dying	(%)	not dying	(%)
Identical	58	57	(98%)	1	(2%)
Non-identical	261	2	(1%)	259	(99%)
Total matches	319	59	(18%)	260	(82%)
Non-matches	2743	27	(1%)	2716	(99%)
Total	3062	86	(3%)	2976	(97%)

2% of the identical matches were incorrect as were 1% of the non-matches.

Table 2 presents the results for the linkage with the 1973 hospital inpatient file containing 600,000 entries. All identical and non-identical matches with a weighting greater than or equal to 220* were examined, a 1 in 3 random sample of those with weights between 170 and 219 and a 1 in 8 random sample of those between 140 and 169. In addition, a 1 in 10 random sample of non-matches was examined. As with the mortality linkage a small percentage (3%) of identical matches were incorrect. Among non-identical matches with weightings less than 220, 5% and 6% were correct and 2% of non-matches were incorrect. These percentages may appear small but the number of discharges they represent form a sizeable proportion of the total discharges.

(b) *Clerical check*

The performance of the clerical check is shown in Tables 3 and 4. Examination of the mortality linkage (Table 3) showed that the clerical check failed to identify the identical match which was incorrect and expressed doubts about one of the fifty-seven identical matches which were correct. Additionally, it failed to identify the two non-identical matches which referred to the same individuals.

The performance of the clerical check in assessing the inpatient linkage (Table 4) shows one of the three identical matches which was incorrect being identified whilst the 114 identical matches were all correctly classified.

*This was the lowest weight at which an identical match appeared.

Table 2. Accuracy of computer linkage for matches obtained with 1973 inpatient file

Type of Match	Number	No. in sample	No. not traced	No. traced and verified as having a discharge	(%)	no discharge	(%)	Estimated No. having a discharge	no discharge
Identical	123	123	6	114	(97%)	3	(3%)	120	3
Non-identical									
weights ⩾ 220	31	31	4	18	(67%)	9	(33%)	21	10
170–219	224	69	13	3	(5%)	53	(95%)	12	212
140–169	348	43	10	2	(6%)	31	(94%)	21	327
Total matches	726	—	—	—		—		174	552
Non-matches	2336	212	30	3	(2%)	179	(98%)	38	2298
Total	3062	—	—	—		—		212	2850

Additionally, the majority of the non-identical matches sampled which were correct were identified. This was achieved without wrongly classifying a large proportion of non-identical matches for which there was no discharge. The clerical check was more reliable for those matches with lower weightings. The loss of reliability at the higher weightings is not critical because a further stage of checking is available in which additional information from hospital case records or other sources can be used to identify those who are wrongly matched.

(c) *Sensitivity and Specificity*

The critical test of the effectiveness of any linkage system is the sensitivity and specificity of the procedure in identifying correctly members of the cohort having subsequent events. Table 5 illustrates for mortality data the levels of sensitivity and specificity obtained using different criteria for accepting a match as correct. By taking only identical matches 66% of deaths reported by the NHSCR tagging scheme were detected. Clerical checking of the limited data did not improve this.

Table 6 illustrates the sensitivity and specificity of the linkage system for inpatient data and shows the number of estimated correct linkages (derived from Table 2, column 6) and the number of estimated incorrect linkages (derived from Table 2, column 7) in the study cohort which would be accepted as correct using the criteria specified. Sensitivity rises from 57% when only identical matches are accepted to 81% when a clerical check is included. This incurs the penalty of including 1.9% of discharges which are incorrect but this is a less serious error as a further stage of checking can eliminate these. It should be noted that the levels of sensitivity obtained in the two lower lines of Table 6 are based on

Table 3. Performance of combined clerical checking in classifying matches correctly; mortality data

Type of match	No. dying	No. classified as dying	% classified correctly	No. not dying	No. classified as not dying	% classified correctly
Identical	57	56	98	1	0	0
Non-identical	2	0	0	259	255	98

Table 4. Performance of combined clerical checking in classifying matches correctly; inpatient data

Type of match	No. having a discharge	No. classified as having a discharge	% classified correctly	No. having no discharge	No. classified as having no discharge	% classified correctly
Identical	114	114	100	3	1	33
Non-identical						
⩾220	18	18	100	9	6	67
170–219	3	2	67	53	50	94
140–169	2	2	100	31	31	100

Table 5. Sensitivity and specificity of mortality linkage

Criteria for acceptance	Sensitivitv No. (%) of the 86 true positives correctly classified as positive	100-Specificity No. (%) of the 2976 true negatives incorrectly classified as positive
Computer listing taking only identical matches	57 (66%)	1 (0.03%)
Computer listing and clerical check	56 (65%)	5 (0.17%)

Table 6. Sensitivity and specificity of inpatient linkage

Criteria for acceptance	Sensitivity No. (%) of the 212 true positives correctly classified as positive	100-Specificity No. (%) of the 2850 true negatives incorrectly classified as positive
Computer listing taking only identical matches	120 (57%)	3 (0.1%)
Computer listing with weighting ⩾220	141 (67%)	13 (0.5%)
Computer listing with clerical check	172 (81%)	54 (1.9%)

sample estimates combined across strata. The figures thus produced are unbiased but are subject to sampling variability.

DISCUSSION

The importance of long term follow-up studies in epidemiological investigations has been recognised for many years. Methods of identifying individuals correctly at points in time which may be well separated are greatly simplified when some constant and easily reproducible information relating to that individual is available. The system of issuing a unique number at birth to all persons in a national population, and providing for all immigrants, is fully established in several countries [7]. In the United Kingdom an identifier is assigned at birth for use in the National Health Service. However, the elements of this identifier are not easily recalled and while, theoretically, it appears on many health documents, in practical terms, its completion is poor [8].

An alternative system, using the surname, initials, date of birth and sex as identifying information for linking, is considered here. The collection of this data routinely on national (Scottish) health documents allows computerised linkage systems to be employed and this paper examines the performance of one such system.

The value of a computerised linkage system depends on a number of factors. Primarily, it should identify a high percentage of true events of interest without including any which are wrongly classified. In addition, work involved in excluding ‘false positives’ should be of manageable proportions. Cost should not exceed that incurred when follow-up is by alternative methods. Finally it is most important that the outcome should be relevant to the investigation being conducted.

This study concentrates on the first of the above requirements and tries to answer the questions—what is the sensitivity and specificity of this linkage system and what additional criteria, if any, can improve these levels?

The use of match weighting and the inclusion of identical matches produces sensitivity levels of 66% for mortality and 57% for inpatient data. Even when accepting all pairings with weights greater than 220 as correct, these levels still remain low for epidemiological investigations. In order to increase these levels the threshold value of the match weighting has to be lowered and this consequently includes a higher proportion of false positives. At present the linkage system can be considered as a screening procedure.

The reasons for low sensitivity are twofold: firstly, mistakes in spelling the surname, initials not agreeing and odd figures out in the date of birth, contribute a large proportion and the majority of these errors can be retrieved in weightings between 140 and 190. Secondly, the required follow-up documents are either not completed or so radically dissimilar that they are missed completely. The linkage system avails little here.

The addition of the clerical check of the limited range of data produced by the linkage system is of considerable value when matching the inpatient data. This contrasts with the linkage using mortality data and is probably because data collected for the inpatient system passes through more stages, each of which may produce errors.

The low sensitivity associated with the mortality data reflects a failure to find the appropriate record in the computer file as compared with the ability to identify and match subjects using the NHS Central Register.

In order to achieve an optimal strategy for this linkage system, a further stage should be added which will virtually eliminate false positives. This requires that the group of matches identified by the clerical checkers as 'probable', be checked against further information either by more detailed examination of the hospital case record or by conventional follow-up.

Thus, levels of sensitivity of 66% for the mortality linkage and 81% for the inpatient linkage have been attained. Specificity can be considered as 100% if the final stage of checking is completed. For the 3062 members of the cohort, this would involve checking 58 (2%) annually for mortality linkage and 183 (6%) for inpatient linkage.

This study has been based on a single linkage system and two sources of data of differing quality. We do not know if this is the best linkage system available or indeed whether it has been adapted to its maximum potential. It may be that this will vary with the characteristics of the files being matched. The ultimate criteria for accepting or rejecting such a system really lies in the answer to the questions—is the sensitivity high enough for the required purpose and is the use of resources to increase the specificity to the desired level of accuracy justified?

For particular purposes, the identifying information may be extended to include other factors such as area of residence and maiden name for females. However, the object of the present study has been to evaluate the system in the most general circumstances.

The extent to which the results reported here can be generalised to other data sets is limited. The levels of sensitivity obtained will vary according to the completeness, accuracy of recording and transcription involved in establishing the data sets. What can be inferred from these results is that in other data sets where the characteristics are similar to those pertaining here and knowledge exists about the relative errors involved, a bounded estimate is available. However, it should be emphasised that very few attempts to validate linkage systems in practical situations have been reported and experience of how they perform with different data sets is limited. Consequently, this paper is more concerned with the description of a methodology for assessing the effectiveness of a linkage system than in the actual results themselves. It is hoped that other linkage systems could be assessed in a similar manner so that a body of data would be available from which comparisons could be made of the way in which data sets with different characteristics influence the sensitivity and specificity of the system.

In summary, outcome in cohort studies can be validly assessed by using computerised record linkage techniques provided certain requirements are satisfied. The sources being used must have a high degree of completeness and, if the accuracy of any of the identifying information is in doubt, a clerical check of the linkage should be incorporated.

Finally, confirmation of the probable matches obtained should be sought using other sources such as hospital case records or death certificates.

Acknowledgements—The authors are grateful to Dr. C. R. Gillis of the Cancer Surveillance Unit of the Greater Glasgow Health Board and to Dr M. A. Heasman and Mrs Vera Carstairs of the Common Services Agency of the Scottish Health Service for their help, criticism and guidance in drafting this report.

The authors are also indebted to Dr J. Dick, Argyll and Clyde Health Board, Hospital Medical Récord Officers and their staffs, the staffs of the Scottish Office Computer Service, the NHS Central Register and to the Provost, Council and citizens of the Burgh of Renfrew for their participation in the study.

The work upon which this report is based, is supported by grants from the Scottish Home and Health Department and the Renfrewshire King Edward Memorial Trust.

REFERENCES

1. Hawthorne VM, Greaves DA, Beevers DG: Blood pressure in a scottish town. **Br Med J** 3: 600–603, 1974
2. Heasman MA: Scottish hospital in-patient statistics. **Med Care** 8 (Suppl): 113–120, 1970
3. Newcombe HB, Kennedy JM, *et al:* Automatic linkage of vital records. *Science* 130: 954, 1959
4. Acheson ED: **Medical Record Linkage.** London, Oxford University Press, 1967.
5. Phillips W: Record linkage for a chronic disease register. In Acheson ED (ed): **Record Linkage in Medicine, Proc. Int. Symp., Oxford, 1967.** Oxford: E. & S. Livingstone, 1968, pp. 120–153.
6. Sunter AB: A statistical approach to record linkage. In Acheson ED (ed): **Record Linkage in Medicine, Proc. Int. Symp., Oxford, 1967.** Oxford: E. & S. Livingstone, 1968, pp. 89–109.
7. Lunde AS: The birth number concept and record linkage. **Am J Public Health** 65: 1165–1169, 1975.
8. Gillis CR: **Ninth Annual Report of the Regional Cancer Committee.** Glasgow: Western Regional Hospital Board, 1971.