A simple, efficient record linkage technic was developed to detect first degree family relationships among hospitalized patients. The procedure was evaluated on the basis of information available from 539 cases of Hashimoto's disease. Potentials and limitations of this epidemiologic tool are discussed.

THE USE OF RECORD LINKAGE TO DETERMINE FAMILIAL OCCURRENCE OF DISEASE FROM HOSPITAL RECORDS (HASHIMOTO'S DISEASE)

Alfonse T. Masi, M.D., Dr.P.H.; Philip E. Sartwell, M.D., M.P.H., F.A.P.H.A.; and Lawrence E. Shulman, M.D., Ph.D.

THE TERM "record linkage" was coined by Dunn in 1946 to designate a process of bringing together two or more separately recorded pieces of information concerning a particular individual or family.[1] Newcombe and his associates developed an electronic process to identify family pedigrees from the vital records of British Columbia.[2-4] Such automatic identification of family units in large populations offers many advantages to epidemiologic and human genetic studies.[5] More recently, record linkage technics have also been applied to inpatient hospital records. For example, in the Oxford record linkage study, all hospital and domiciliary admissions as well as vital events occurring after January 1, 1962, are being compiled for a defined population of 325,000 persons.[6] Also, in Aberdeen, Scotland, a psychiatric case register is being developed and record linkage is being used to identify individuals and families who appear in the accumulated records.[7] Thus, the application of record linkage to vital records is hardly older than one decade and its application to hospital morbidity data is in its infancy.

Materials and Methods

In this study, record linkage was applied to hospital records of patients with Hashimoto's disease or chronic thyroiditis who were detected in a survey of all general hospitals in Baltimore. These two diagnoses were found to be interchangeable and hereafter we will use the term Hashimoto's disease. Hashimoto's disease is a thyroid disorder which is usually present as a goiter or a nodule but which can only be diagnosed with assurance, in most cases, by its pathological appearance. The primary purpose of our epidemiologic survey was to determine the frequency and distribution of this disease in a defined population. As a secondary objective, we used record linkage to investigate the degree of familial occurrence among the cases detected in this survey because some controversy exists concerning a familial predisposition.

Data were collected from all 17 general hospitals of Baltimore for the period from 1948 to 1960 inclusive. Cases were limited to those identified in hospitals by means of surgery or biopsy and who had a histological diagnosis of Hashimoto's disease or chronic thyroiditis. Record linkage was performed with personal identifying information available on the hospital admission sheet. In some instances this sheet was not available because the original hospital record was not obtained. Such cases were included in the survey on the basis of a clinical summary and a pathological report.

First degree relationships, namely full sibling and parent-child, were identified by a mechanical process of association by name. A standard 80-column IBM card was used to enter specific names. Each letter of a name can be recorded in one column of the card. The family name or surname was entered first and one space was allowed before entering given names or initials. Three fields of the card, each of 20 columns, were used for entering names. One free space was reserved between each field. Entire names were transcribed without coding, except for unusually long names, in which case the given name was abbreviated. The names entered into the three fields were as follows: the current name of the case; the natural father's birth name; and lastly the natural mother's birth or maiden name. A last field of 17 columns was provided for data to identify the case (Figure 1).

After the cards were sorted in alphabetic order in a particular field by a standard IBM sorting machine, all the information on these cards was reproduced by a mechanical printer. Family linkage was then looked for by visually scanning the alphabetized name lists for duplicated names.

Full-sibling linkages may be found by sorting alphabetically the case identification cards on either of the two fields devoted to the fathers' or mothers' birth names. Duplication of the name of both parents indicates a full-sibling relationship. Half-sibling relationships may be revealed by separately sorting on both the column sections devoted to fathers' and mothers' birth names.

Father-child relationships may be identified among the cases because each has a father with the same family name or surname. Therefore, such linkages may be found by sorting alphabetically the case identification cards in the field devoted to fathers' birth names. Duplicated family names may then be identified visually. Given or Christian names in this list may vary since we are not linking the name of a single individual, such as a parent in the sibling linkage, but are only linking a common family name. Duplication of the full name of the father who is a case does occur, but in separate fields. That is, since the father is a case, his full name occurs once in the current name field. It appears again in the father's birth name field of his child who is also a case. These duplicated full names should be in adjacent rows.

Mother-child relationships are best identified by using the birth name or maiden name of the mother. It should be noted, however, that the birth name of a case is not one of the three names entered in the primary case identification card. In order to associate mother-child relationships by using the birth or maiden name of the mother, an auxiliary identification card is also made for each female who has ever been married (Figure 1). The assumption is made that any female who has ever been married may also be the mother of another case in the series. The auxiliary card is identical to the primary card, except that the field, which on the primary card contains the mother's birth name, in the auxiliary card contains the birth name of the case. Mother-child relationships may now be identified by

sorting alphabetically all primary and auxiliary cards in the maiden name field. If a mother-child relationship exists, then the mother's own maiden name will be next to the mother's maiden name of her child. Figure 2 illustrates how an actual mother-child linkage was identified among the cases in this series. This was discovered even though the mother was twice married and her daughter was also married.

Results

In this survey of Baltimore hospitals, 539 cases of Hashimoto's disease were discovered. It was diagnosed in 9.1 per cent of thyroid operations on females and in 2.6 per cent of thyroid operations on males. A reasonably high proportion of the hospital records had parental birth name information (Table 1). Forty-one cases (7.6 per cent) had no parental name information because, in most of these incomplete records, the hospital admission sheet was not available. Sibling relationships could potentially be obtained only among the 498 cases where either the father's or mother's birth name was given. Father-child relationships could potentially be obtained among the 496 cases where the father's birth name was available. Mother-child relationships could be obtained only among the 309 cases where the mother's birth name was given.

Spurious linkage refers to the coincidental association of names which does not reflect a familial relationship. Incomplete knowledge of parents' birth names will, of course, make this linkage process less efficient. For example, 19 instances of coincidental linkage occurred on incomplete names in this

Figure 1—Schematic IBM Case Identification Cards Used in Identifying Familial Relationships by Name Association (Record-Linkage)

1. Primary case identification card*

SUGGESTED FIELDS FOR KEY PUNCHING CASE IDENTIFICATION DATA			
1 20	22 41	43 62	64 80
PATIENT'S CURRENT NAME	NATURAL FATHER'S BIRTH NAME	NATURAL MOTHER'S MAIDEN NAME	CASE IDENTIFICATION DATA

* This is the only card required for males or for females who never changed their maiden name.

2. Auxiliary case identification card*

SUGGESTED FIELDS FOR KEY PUNCHING CASE IDENTIFICATION DATA			
1 20	22 41	43 62	64 80
PATIENT'S CURRENT NAME	NATURAL FATHER'S BIRTH NAME	PATIENT'S OWN MAIDEN NAME	CASE IDENTIFICATION DATA

* This card is prepared for all adult females unless it is known that they have never changed their maiden name.
N.B. Only field 43 to 62 differs between cards 1 and 2.

Figure 2a—Hypothetical Family Pedigree with a Mother and a Daughter Having Hashimoto's Disease

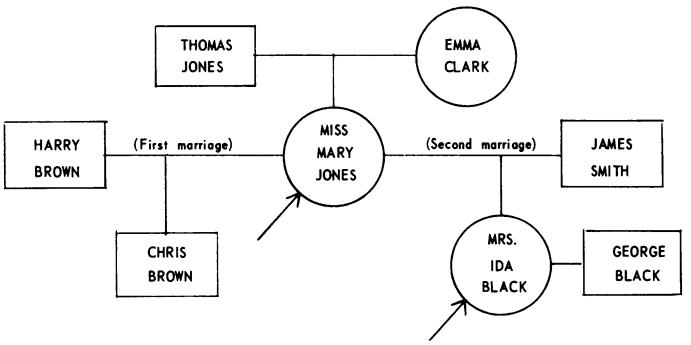

Figure 2b—A Schematic IBM Identification Card and Print Out of Cases Sorted Alphabetically on Maiden Names (Note the Association of a Mother and Her Daughter)

SUGGESTED FIELDS FOR KEY-PUNCHING CASE IDENTIFICATION DATA				
1 – 20	22 – 41	43 – 62	64 – VARIABLE – 80	
PATIENT'S NAME	NATURAL FATHER'S BIRTH NAME	PATIENT'S AND MOTHER'S BIRTH NAME	SERIAL NUMBERS HOSPITAL	PATIENT
		ALPHABETICALLY SORTED IN THIS FIELD		
SMITH, MARY	JONES, THOMAS	CLARK, EMMA*	03	026
MASON, NANCY	KEY, EMIL	JOHNSON, LUCY	17	037
SMITH, MARY	JONES, THOMAS	JONES, MARY **	03	026
BLACK, IDA	SMITH, JAMES	JONES, MARY	09	021
GRAY, BARBARA	MILLER, HENRY	JORDAN, ANNA	06	103
ABBOTT, SUSAN	GREEN, CHARLES	ROBERTS, CATHERINE	07	008

* Mrs. Mary Smith's mother's maiden name.
** Mrs. Mary Smith's own maiden name.

study. The given name of the father and the birth name of the mother were not available in these instances. Reference to age and race information in the original records excluded eight of these associations. Such data could also have been added to the IBM cards in the field containing identifying information. In the remaining 11 instances a telephone call to the physcians attending the cases provided the missing parental name information. There were no true linkages among the 19 associated incomplete parental birth names.

Five confirmed first degree family relationships were found by this linkage process. Four families had two or more first degree relatives with a reported pathologic diagnosis of Hashimoto's disease or chronic thyroiditis (Table 2). In the fifth family, one of two sisters was listed in a pathology diagnostic index as having chronic thyroiditis; but her pathology report was "scarring and a rare focus of lymphoid infiltration." Her sister had early Hashimoto's disease at her second thyroid operation. This number of families with multiple first degree relatives having histologically diagnosed Hashimoto's disease is relatively large in terms of the total reports which we found in the English language literature. Only five such families were previously reported, in-

Table 1—The Number and Per cent of Cases Who Had Parental Family Names Specified on Available Hospital Records, Baltimore, 1948-1960

Parental Birth Names	Number	Per cent
Father's and mother's*	326	60.5
Father's only	170	31.5
Mother's only	2	0.4
Neither	41	7.6
Total cases	539	100.0

* In 19 only the mother's given name was provided.

cluding a set of uniovular female twins.[8-12] To our knowledge, the record-linkage process identified the only father-child relationship with histological confirmation.

Discussion

Record linkage in this study required little time or expense. The primary case identification cards were punched and verified in ten hours; auxiliary cards required less time, since most of the information was mechanically reproduced from the primary cards. Less than two hours was required to sort the cards alphabetically on both of the parental birth name fields and to print out the information mechanically. Visual inspection of the print-outs was done on two separate occasions by two observers in eight person-hours. It is difficult to say at what point the size of a study would exceed the limits of this procedure and would require conversion to an electronic process. In a larger study where increased discriminating power of the identifying information is needed due to larger files of names, this can be obtained by supplementing the name information with other particulars such as age, race, religion, and place of birth. Where the best information has been put on the cards for the purpose of visual linkage, this will probably also be best for the purpose of computer linkage.

Regarding the application of record linkage to hospital data in general, two broad classes of hospital statistics have

Table 2—Families in Which Two or More First-Degree Relatives Had Hashimoto's Disease or Chronic Thyroiditis

Family	Relationship	Age at Operation	Date of Operation	Pathologic Diagnosis	Hospital to Which Admitted for Surgery
I	Daughter	40	12–10–56	Colloid adenoma and Hashimoto's disease	13
	Father	71	7–10–57	Chronic thyroiditis	07
II	Mother	48	1–7–54	Hashimoto's disease	17
	Daughter	24	2–23–57	Hashimoto's disease	17
III	Sister 1	27	4–20–56	Chronic thyroiditis	03
	Sister 2	29	5–28–56	Hashimoto's disease	03
	Sister 3	31	1–21–58	Chronic thyroiditis	03
IV	Sister 1	50	2–28–50	Hashimoto's disease	13
	Sister 2	53	1–9–56	Early Hashimoto's disease	13
V	Sister 1	46	5–14–54	Early Hashimoto's disease	13
	Sister 2	50	2–20–57	Scarring and a rare focus of lymphoid infiltration*	13

* Listed in the Pathology Department index as chronic thyroiditis.

been outlined by the Expert Subcommittee on Hospital Statistics of the World Health Organization: (1) selected hospitals without possibility of measuring the population at risk; and (2) all hospitals, or special hospitals, in a defined geographic area where diagnoses may be placed in relation to the population at risk.[13]

As examples of the former class of hospital morbidity data, we would like to cite two reported investigations which could have benefited from automated record linkage processes. In 1960 Humphries reported a notable study of the occurrence of hypertensive toxemia of pregnancy in mother-daughter pairs in which both had babies delivered on the Johns Hopkins obstetrical service.[14] A full year was required to derive manually the 300 mother-daughter pairs by name association. Approximately 8,000 obstetrical records of daughters were checked against old obstetrical service discharge files in search of possible mothers. The tedium of this valuable investigation could have been reduced markedly by record linkage.

The second example of a selected hospital study which could have been assisted by record linkage was reported by Chesley and his associates in 1962.[15] They studied pregnancies in sisters of 107 women with eclampsia, all of whom were delivered at the Margaret Hague Maternity Hospital. The sisters were identified by questioning the eclamptic women who were under long-term observation. They also traced daughters of eclamptic women delivered at more than 40 different hospitals. Record linkage alone without follow-up or interview could have identified objectively the sisters delivered at this hospital. It also could have provided additional sisters of noneclamptic index women to serve as controls in this investigation.

Applications of record linkage to the second class of hospital data, data from a defined "hospital service area," are even more valuable.[16,17] Hospital case records may be important sources of morbidity statistics, especially for rare diseases and for those requiring hospitalization.[16] On the other hand, hospital morbidity data may be highly selective and not representative of morbidity in the community.[13,18,19]

It is probable that our city-wide survey of Hashimoto's disease based upon hospitalized cases does not depict with complete accuracy its distribution in the population. The method of diagnosis, that is by pathologic findings, necessitates this type of survey. The question to be asked, therefore, is whether any useful knowledge can be gained from linking hospital diagnosed cases into families. This question must be considered in the light of available information on the familial occurrence of this disease. In the scattered reports of five families which we found, no definition of the population at risk to this disease and no estimate of the risk were available.[8-12] It is not surprising that conflicting conclusions have been drawn regarding a familial predisposition to Hashimoto's disease from such reported familial groupings unrelated to any defined population.[10,20] Consequently, community-wide hospital survey data may contribute useful knowledge because a population and its risk to disease may be defined at least crudely.

Regarding selection bias with respect to hospitalization and diagnosis in our survey, there are factors which might tend to exaggerate any possible familial aggregation: (1) because Hashimoto's disease is relatively asymptomatic, there is probably an enhanced likelihood of detecting a second case in the family after an initial case has been discovered; (2) socioeconomic factors influencing the admission of patients to hospitals might also tend to exaggerate any familial aggregation.

On the other hand, there are factors which might underestimate the degree

of familial occurrence, such as: (1) incomplete enumeration of cases, and (2) migration of affected relatives out of the Baltimore "hospital service area." After careful consideration, we believe it more likely that the forces tending to exaggerate familial aggregation in this survey outweigh those tending to underestimate it.

Thus, if no greater familial occurrence of disease among hospitalized cases is found than one might expect from the risk of disease in the general population, then it is reasonable to doubt that any familial aggregation exists. On the other hand, if there is a striking excess, such as 10- or 20-fold, in the observed familial occurrence of disease among hospitalized cases compared with an expectation based on population risks, then it is reasonable to believe that familial aggregation exists. This would be especially true in populations where hospitalization is generally available and in diseases for which hospitalization is usually needed.

If, however, the results fall between the above extremes, then a decision as to whether familial aggregation occurs will be difficult on the basis of hospital morbidity data alone. Such was the situation in this study. All members of the five multiple-case families were residents of urban Baltimore. This allowed increased comparability with a geographically defined population.[19] Since the survey extended over a relatively restricted time interval of only 13 years, the degree of parent-child aggregation was not analyzed. It is assumed that an average calendar interval of a generation or approximately 25 years might occur between diagnoses of Hashimoto's disease in parents and their children under stable circumstances of medical detection. Only the degree of sibling aggregation was analyzed. It is apparent that except when a survey interval embraces the entire life-span of all relatives, detection of multiple cases in families will be incomplete from any record source. The shorter the interval of a survey the less complete it will be. It will be much less complete for parent-child relationships than for sibling relationships.

These data are not presented as establishing family aggregation of Hashimoto's disease. Many considerations which cannot be discussed here are involved in the interpretation of these data. Based on age-adjusted Baltimore white female risks of Hashimoto's disease as determined from this hospital survey, one sister-sister relationship might be expected among the 262 Baltimore white female cases instead of the three observed families.[21] The selection factors cited above might account for the difference but some true familial aggregation cannot be excluded.

We believe that record linkage can be a valuable first step in investigating familial aggregation of disease when appropriate data are available. It is especially valuable when the disease under study requires hospitalization; when population risks can be estimated; when the survey interval is sufficiently long; when the disease is infrequent in the population; and when a hospital service area, serving a stable population, can be utilized. Of course, finding unequivocal family aggregation does not necessarily connote a genetic basis for a disease but may reflect environmental factors such as common exposure to an agent of disease. The definitive determination of family aggregation will continue to rest upon the investigation of families of affected index cases in comparison with the families of index control individuals.[22]

Summary

A simple and efficient record-linkage technic was developed to detect first degree family relationships among hospitalized patients. This procedure was

assayed with information available in the records of 539 cases of Hashimoto's disease discovered in a city-wide hospital survey. Five new families were discovered in which this condition was pathologically diagnosed in two or three immediate relatives. Some potentials and limitations of this relatively new epidemiologic tool as applied to hospital records are discussed.

ACKNOWLEDGMENTS — The authors are indebted to personnel at each of the general hospitals in Baltimore for their generous cooperation in this study. It is not possible to mention by name each of the many persons who helped to obtain necessary information in this survey. Special thanks are offered to Mrs. Norma Lutins of the Department of Biostatistics of the Johns Hopkins School of Hygiene and Public Health for her generous technical assistance in processing these data.

REFERENCES

1. Dunn, H. L. Record Linkage. A.J.P.H. 36:1412–1416 (Dec.), 1946.
2. Newcombe, H. B. Detection of Genetic Trends in Public Health *in* Effect of Radiation on Human Heredity. Geneva, Switzerland: World Health Organization, 1957, pp. 157–168.
3. Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. Automatic Linkage of Vital Records. Science 130:954–959 (Oct. 16), 1959.
4. Newcombe, H. B. Chapter XI—Feasibility of Estimating Consequences of an Increased Mutation Rate *in* Molecular Genetics and Human Disease. Gardner, L. I. et al. Springfield, Ill.: Thomas, 1960, pp. 186-202.
5. The Use of Vital and Health Statistics for Genetic and Radiation Studies. New York, N. Y.: United Nations, World Health Organization, 1962, 259 pp.
6. Acheson, E. D. Identification of Documents in the Oxford Record Linkage Study (Abstract). Brit. J. Prev. and Social Med. 17:46 (Jan.), 1963.
7. Baldwin, J. A. Plans for a Psychiatric Case Register for Aberdeen, Scotland. Seminar held in the Department of Epidemiology, The Johns Hopkins University School of Hygiene and Public Health, November 13, 1963 (not in publication).
8. Craig, P. E.; Spann, J. L.; and Lowbeer, L. Hashimoto's Disease (Struma Lymphomatosa)—A Familial Incidence of Three Cases. Am. J. Surg. 84: 286–292 (Sept.), 1952.
9. Dunning, E. J. Struma Lymphomatosa: Report of Three cases in One Family. J. Clin. Endocrinol. 19: 1121–1125 (Sept.), 1959.
10. Irvine, W. J.; MacGregor, A. G.; Stuart, A. E.; and Hall, G. H. Hashimoto's Disease in Uniovular Twins. Lancet 2:850–853 (Oct. 14), 1961.
11. DeGroot, L. J.; Hall, R.; McDermott, W. V.; and Davis, A. M. Hashimoto's Thyroiditis—A Genetically Conditioned Disease. New England J. Med. 267: 267–273 (Aug. 9), 1962.
12. Hung, W., and Winship, T. Struma Lymphomatosa in Mother and Daughter with Serologic and Histologic Evidence. J. Clin. Endocrinol. 23:465–469 (May), 1963.
13. Expert Committee on Health Statistics. Report on the Second Session. Wld. Hlth. Org. Tech. Rep. Ser., No. 25, Geneva, 1950.
14. Humphries, J. O. Occurrence of Hypertensive Toxemia of Pregnancy in Mother-Daughter Pairs. Bull. Johns Hopkins Hosp. 107:271–277 (Nov.), 1960.
15. Chesley, L. C.; Cosgrove, R. A.; and Annitto, J. E. Pregnancies in the Sisters and Daughters of Eclamptic Women. Obst. & Gynec. 20:39–46 (July), 1962.
16. Mooney, H. W. Methodology in Two California Health Surveys. Pub. Health Monogr. No. 70, PHS. Publ. No. 942. Washington, D. C.: Gov. Ptg. Office, 1962.
17. Hess, I.; Riedel, D. C.; and Fitzpatrick, T. B. Probability Sampling of Hospitals and Patients. Ann Arbor, Mich.: University of Michigan, 1961.
18. Expert Committee on Health Statistics—Fifth Report. Wld. Hlth. Org. Tech. Rep. Ser., No. 133, Geneva, 1957.
19. Expert Committee on Health Statistics—Third Report. Ibid. Ser., No. 53, Geneva, 1952.
20. Danowski, T. S. Clinical Endocrinology. Vol. 2, Chapter Thyroiditis. Baltimore, Md.: Williams and Wilkins, 1962, pp. 429-453.
21. Masi, A. T. Hashimoto's Disease: An Epidemiological Study Based on a Community-Wide Hospital Survey. (To be published in J. Chronic Dis.)
22. Haenszel, W. Some Problems in the Estimation of Family Risks of Disease. J. Nat. Cancer Inst. 23: 487–505 (Sept.), 1959.

Dr. Masi is assistant professor of epidemiology, and Dr. Sartwell is professor of epidemiology and chairman, Department of Epidemiology, The Johns Hopkins University School of Hygiene and Public Health. Dr. Shulman is assistant professor of medicine, The Johns Hopkins University School of Medicine, Baltimore, Md.

This paper was presented before a Joint Session of the American School Health Association and Epidemiology Section of the American Public Health Association at the Ninety-First Annual Meeting in Kansas City, Mo., November 12, 1963.

This research was initiated under a training grant No. 2G-20 provided by the National Institutes of Health, while the senior author was a Doctor of Public Health candidate and a post-doctoral trainee in epidemiology at the Johns Hopkins School of Hygiene and Public Health. This study was a part of Dr. Masi's doctoral thesis, accepted in May, 1963. This research was also supported in part by a Public Health Service research grant NIAMD No. AM 06726, and by a Public Health Service special fellowship 1F3 AM-21, 397-01.