Sozial- und Präventivmedizin Médecine sociale et préventive 30, 68 – 71 (1985)

Integrating Health-Related Data from Various Sources: Combining Surveys, Records and Routine Data*

Elisabeth Schach

University of Dortmund,
Postfach 500 500, 4600 Dortmund 50, F.R.G.

INTRODUCTION

Integrating data from various sources is desirable in any area characterized by multiple data use. In health services administration and research, in descriptive and analytical epidemiological studies, and in routine health statistical collection activities, data collection is often costly and/or time consuming. Due to this, such data sets should not only be used for the single purpose they were collected for, but they should also have the potential to be put to multiple use. Thus, they should fill the pre-requisites so as to be used as secondary data by others in conjunction with further relevant data sets. In particular, large samples or population censuses performed with a similar methodology for a long time would gain in usefulness, if this would hold true for them.

Monitoring the health of populations, describing trends in mortality and morbidity, or describing the use of health services require not only a variety of different data sets but also imply the necessity of combining data in a way that is adequate for the questions being asked.

As not all potential data uses for a particular data set may be forseen in advance, ongoing, particularly large routine data sets may become potentially useful if several requirements for them are met. In countries without national personal identification numbers or similar schemes, the existence of such formal characteristics is a necessary condition for any secondary data use. These formal criteria involve methodological characteristics, which will be discussed below. Discussing the issue of integrating data from various sources may be done from several perspectives, namely:

- linking data at one level of aggregation
- linking different data components at one aggregation level
- linking multiple data components at several levels of aggregation

LINKING DATA AT ONE LEVEL OF AGGREGATION

In health services research, it is useful to link data at an individual level or for groups of persons (i.e. aggregates of persons). Much practical experience is already available for performing data linkage at an individual level. It was developed by computer specialists and is based on linking data from several data sets using one or more characteristics common to both sets. This linkage may be performed on a deterministic or a probabilistic basis. It is routine with health or life insurance companies. Data are linked by insuree and provider in such settings.

When linking data for groups of individuals, characteristics of these groups have to be common to both data sets to be linked. Often, locality, area or scheme number may be used as linkage parameters. If the data set's locality or area definitions agree with definitions of geographic sub-areas of the national statistical system, then useful demographic information may be drawn from that system.

Obviously, patient data may be linked to physician data if service or contact data sets for patients routinely contain the physician number. On the other hand, separate patient data components may be combined if services data sets contain reliable and complete patient identification information, such as name, birth date, place of birth or insurance number.

In spite of these possibilities, linking data of public health services schemes and other sources, or health services research on the basis of linked data are still relatively uncommon in some countries. There are many possible reasons for this. They are :

- unclear ideas about what use integrated data sets may be put to
- payment or reimbursement schemes may not require such linkage. In fee-for-service schemes, for example, documents may contain just types and dates of delivery services, which are required as the basis for remuneration (i.e. there is no administrative requirement for linkage)
- fear of the cost and effort required to handle large volumes of data as preparation for producing linked data sets
- fear of misuse of the linked data sets
- fear of the effort required to introduce uniform standards and definitions as a pre-requisite for linking data sets.

As a consequence of this separateness of data, we find much duplication of data collection activities, particularly of random samples attempting to collect data which exist in census form elsewhere. As a matter of fact, in the F.R.G., collecting data of persons using physician services in ambulatory care via random samples of physicians or the population seems much simpler than drawing samples of documents about the population's contacts with physicians and deducing the information in this way (note that in a fee-for-service system as, for example the F.R.G., data about almost each patient-physician contact and the services delivered are stored for reimbursement purposes). Even though each service act delivered in ambulatory care in the country is entered into a computer, reliable estimates of the number of blood pressure checks per 1000 population, or per 1000 persons of a specified age group are hard to produce for a given time period.

This demonstrates that linking of data at various levels of aggregation means that a number of formal requirements be met, namely, that reliable patient and complete provider identification or characteristics be available and agree in all data sets to be linked at an individual level. The more formal requirements are the following :

- The population references (e.g. sub-populations, members of a scheme) are available and identical in all data sets to be linked at an aggregate level. The population should be a defined population.

* Paper prepared on the basis of a presentation given at the Xth scientific Meeting of the International Epidemiological Association, Vancouver, B.C., 19-25 August 1984.

- The reference to persons should be possible or producible. These persons might be patients or providers, depending on who the linkage is planned for.
- The period reference be available and identical for all data sets to be linked at an individual or aggregate level.
- The procedure definitions are available and identical for all data sets to be linked, if this should be the linkage criterion.
- The place or geographic subdivisions should be identical and available for all data sets to be linked. This is essential for linkage at any area level, and particularly critical at a smallarea level.
- Other parameters, terms, and definitions of variables should agree across data sets, e.g. age, occupation, desease categories, if occupational health studies are under consideration.
- The classification schemes must agree across data sets if grouped data are to be used for analysis or linkage. This means that identical subdivisions are required for variables, such as age or diseases, if linking is performed for such groups.

LINKAGE OF DIFFERENT DATA COMPONENTS AT ONE AGGREGATION LEVEL

The production of data sets covering several content areas may further be achieved by collecting data at a particular aggregation level about a variety of topics. Many social surveys use this approach by collecting data from various content areas with one survey instrument. The areas that data collection focuses on may be vital and economic data, occupational information, health services use, morbidity, etc. Such survey procedures are applied in one-time data collection or ongoing data collection activities. Examples of social surveys using this integrated approach are the general household survey of England (1), The Mikrozensus of the Federal Republic of Germany (2) and many other health and social surveys. Both surveys have been repeated several times so that trend information about the various subject areas, including health, may be derived.

The limitation of many of these surveys has been that the resulting data sets were often neither comparable across various data collectors nor over time within the same country. This lack of continuity limits the role surveys could play for ascertaining the health of populations (3). International comparisons are hardly possible under these circumstances. Thus, suggestions have been made in our country to agree on uniform data collection techniques, terms and definitions for a data component required in almost any population field survey and many routine data sets: namely the vital statistics portion. All major commercial survey organizations agreed to use that uniform vital data component. Thus, data from different surveys may potentially be linked or compared on that basis.

Linkage of different data components may also be possible via the minimum basic data sets developed in the U.S.A. for hospital (4) and long-term (5) care. The uniformity in these minimum basic data sets goes beyond the uniformity described in the previous paragraph. The minimum basic data sets cover several subject areas, such as socio-demographic, health services category, disease and procedure information in a uniform way across persons, institutions and providers and thus tremendously improve the analysis and linkage potential of data sets from various settings and of different contents, given only that the minimum basic data sets agree and are used by many settings.

A somewhat more universal instrument of linking different data sets at one aggregation level is provided by the United Nation's Health Survey Capability Program (6). The merit of that program consists in its potential to compare similar data sets internationally without the need to organize international survey programs at the same time. This is achieved by recommending to participating countries standardized data collection, instruments, and methods for field surveys in a modular arrangement. This arrangement provides for the freedom of countries to choose their topics of interest and then to use the relevant modules of the program in order to implement the necessary data collection (6,7). It covers topics such as nutrition, health, morbidity, special client groups, and the use of the health services (7).

LINKING SEVERAL DATA COMPONENTS AT SEVERAL LEVELS OF AGGREGATION

When attempting to link multiple data components at varying aggregation levels, then the formal requirements discussed above must be given for respective pairs of data sets at the respective linkage levels. This means that when it is intended to link patient data to their respective physicians, patient as well as physician data sets must contain identical physician identifiers. When such linked data is then to be related to demographic data from the official statistical system in order to estimate rates of use or ratios of physicians per population, then the geographic areas for which this is planned have to be identical in both data sets.

On the other hand, if environmental risks are to be related to rates of specific morbidity, then morbidity information has to be ascertained for the areas with environmental risk measures. So it is evident that, before planning major studies, it would be wise to ascertain which other relevant data sources might be useful for the investigation, and how one might be able to tap that data within the study's frame of reference.

EXAMPLES OF STUDIES BASED ON MORE THAN ONE DATA SET FROM DIFFERENT SOURCES

In the following section three examples are given demonstrating the combination of data from different sources. In epidemiological research, the need to combine data in such a way comes about from the need to answer questions about possible associations between several variables. This might be a risk factor (air pollution) and an outcome variable (rate of hospitalizations for croup) as in example 2. We might also be interested in studying an outcome variable, such as cancer incidence and relate it to a process variable, such as smear-taking as in example 1. Both of these studies performed the linking after the original data collection activity had been completed, but data linkage at various aggregation levels may also be part of a study design, as is demonstrated in example 3.

1. Regional trends in incidence of cervical cancer in Denmark in relation to local smear-taking activity (8)

The investigation studies the relationship between trends in the incidence of cervical cancer in Denmark between 1943 and 1977 and pap-smear taking activities. For that purpose, data from the Danish National Health Insurance scheme, the organized screening programs, hospitals, and the national cancer registry are combined. "The study shows a considerable decline in the cumulative incidence of cervical cancer for women aged 30-59 years of age in areas where organized screening programs began before 1969. The decline in areas without organized programs but with a higher or equivalent screening activity in 1974/75 is in general smaller or occurring later"(8).

Summary

Study purpose

Study of the relationship between trends in cervical cancer incidence and pap-smear taking.

Data sources

- Danish national cancer registry for cervical cancer incidence

- National Health Insurance scheme for pap-smear taking outside of organized screening programs
- Hospital admissions for gynecological reasons for pap-smear taking in hospitals
- Organized screening programs for pap-smear taking within those programs
- Official statistics system for demographic data

Statistics produced

- Rates of cumulative cervical cancer incidence per county of Denmark (5-yr periods)
- Rates of pap-smears per woman aged 30 - 59 per county of Denmark in 1974/75.

Linkage level

- County (old or new) of Denmark

Difficulties

- Change of county borders during study period.

2. Relationship between air quality and respiratory disease among children aged 0 - 4 years (9)

The study relates the incidence rates of the croup syndrome and of obstructive bronchitis in children aged 0 - 4 years to the level of air and dust pollution in the areas where the children live. The authors show a positive correlation between the incidence rates of the croup syndrome and of bronchitis and the level of SO_2 and dust pollution in small geographic regions after adjusting for several intervening factors, such as distance to clinic, level of infection rate (9).

Summary

Study purpose

Study of relationship between disease incidence rates in children and air pollution level.

Data sources

- Hospital admissions in a 3 year period for the croup syndrome and for bronchitis in children as a basis for incidence rate calculation
- SO_2 emissions and dust emissions for 60 one-km^2 areas for one-quarter periods

Statistics produced

- SO_2 level in mg per m^2, dust level in g per m^2*d
- Admission rates to one children's hospital for the croup syndrome and for obstructive bronchitis in the age group 0 - 4 years.

Linkage level

- One-km^2 areas

Difficulties

- Possible selectivity of cases to one hospital
- Area definitions somewhat artificial

3. Survey among ambulatory care physicians in the Federal Republic of Germany (EVaS-Study) (10)

The survey was carried out among a random sample of physicians who documented a sample of their patient contacts with their offices. The purpose of the study was to learn about the contents of ambulatory medical care in the Federal Republic of Germany and several of its sub-regions by recording reasons for contact, patient characteristics, services delivered and recommended, diagnoses and disposition. Physician characteristics were ascertained by a mail survey.

Summary

Purpose

Study of contents of ambulatory medical care in the F.R.G. in 1981/82 and some of its sub-divisions

Data sources

- Office and personal patient-physician contacts from a sample survey in physicians' offices
- Office characteristics and resources from a mail questionnaire
- Physician characteristics from a national register of physicians
- Socio-demographic data about studied regions from official statistics

Statistics produced

- Volume of physician contacts per 1000 population for the whole area and its subdivisions and by patient characteristics
- Ratio of physicians per 1000 population by speciality for the total area and its subdivisions
- Response rates and non-respondents' characteristics

Linkage levels

- Contacts with physicians by physicians
- Contacts with physicians by area
- Physician survey with physician registry data
- Contacts with physicians with regional demographic data

Difficulties

- Lack of availability of data about foreigners by regions
- Lack of data about members in health insurance schemes by regions

These three examples from very different research areas show that linking of data sets from different sources is frequently performed on the basis of regions as linkage parameters. Knowing this, geographic information in data sets should be recorded in a uniform way. It should, for example, be recommended that geographic information be stored in the form of standard statistical areas or postal zip codes, if these have reached a certain stability over time. Such standardization would obviously greatly enhance the usefulness of routine data sets for small area analyses.

Summary

The paper discusses the necessity of combining data from various sources in order to enhance their usefulness for a variety of applications. As future data use can hardly be forseen in advance, major data sets in health services should fulfil several formal requirements in order to make them suitable for future linkage. These formal requirements are that there be references to defined populations, to specific persons, to defined time periods, to specific places or regions. It would be necessary for terms, definition and classification schemes to agree between data sets which are to be linked and be in wide use. Three facets of data linkage are discussed specifically namely linking data at one level of aggregation, linking different data components, and combining data sets from different sources at several levels of aggregation. Three examples are provided, describing linkages of data from various sources for epidemiological studies and a study in health services research. They show that at this point in descriptive epidemiological studies linkage on the basis of regions is of great importance. This implies that it would be desirable for large scale data collection activities in health services to provide for a uniform representation of the geographic areas. Such uniformity would greatly enhance the linkage potential of data sets and thus their usefulness for small area and regional analyses.

Zusammenfassung

Verknüpfung von Datensätzen im Gesundheitswesen

Es wird die Notwendigkeit diskutiert, Daten aus verschiedenen Quellen zu kombinieren, um deren Nutzen für unterschiedliche Anwendungen zu erhöhen. Da zukünftige Datennutzung für einen Datensatz nicht vorausgesehen werden kann, sollten wichtige Datensätze im Gesundheitswesen einigen formalen Kriterien genügen, um sie für zukünftige Verknüpfungen geeignet zu machen. Diese formalen Kriterien sind der Bezug zu einer definierten Population, Personenbezug (es kann sich um Patienten-und/oder Versorgerbezug handeln), Bezug zu einer definierten Zeitperiode und zu definierten Orten oder Regionen. Ausserdem müssen in zu verknüpfenden Datensätzen Termini und Definitionen, sowie Klassifikationsschemata übereinstimmen. Datenverknüpfung wird für drei Situationen diskutiert, nämlich die Verknüpfung von Daten auf einer Aggregationsstufe, Verknüpfung von Datenkomponenten und die Verknüpfung unterschiedlicher Datenkörper auf unterschiedlichen Aggregationsstufen. Es werden drei Beispielen beschrieben, die Datenverknüpfungen für epidemologische Studien und eine Studie im Gesundheitswesen zeigen. Aus diesen Beispielen geht hervor, dass der Verknüpfung von Daten auf der Basis von Regionen heute eine grosse Bedeutung zukommt. Daraus folgt, dass es wünschenswert wäre, in wichtigen Datenkörpern des Gesundheitswesens die Variable Region in einheitlicher Weise darzustellen. Dies würde deren Brauchbarkeit über den eigentlichen Zweck der Datenerhebung hinaus, insbesondere aber für Kleinräumige Analysen, erheblich erhöhen.

Résumé

Fusion de diverses sources de données issues des service de santé

Cet article analyse le potentiel de l'utilisation de diverses sources de données provenant du système de santé. Pour permettre une utilisation pour des buts divers, et à plusieurs reprises, sont developpés une série de critères auxquels doivent satisfaire les banques de données. Trois exemples montrant l'intérêt de la fusion de plusieurs sources dans des études épidémiologiques où la recherche sur les services de santé sont discutés.

References

1. Office of Population Censuses and Surveys (OPCS). Social Survey Division. The General Household Survey. Introductory Report. London: Her Majesty's Stationary Office, 1973.
2. Brennecke R. Mikrozensuserhebung. In: Brennecke R., Paul H.A., Greiser E. und Schach E. (Hrsg.) Datenquellen für Sozialmedizin und Epidemiologie. Heidelberg: Springer 1981.
3. Schach E. Die Messung des Gesundheitszustands der Bevölkerung : Rolle und Güte von Befragungsdaten. Sozial-und Präventivmedizin No. 2 p. 65-71.
4. National Center for Health Statistics (NCHS). Uniform Hospital Discharge Data. Minimum Data Set. Report of the National Committee on Vital and Health Statistics. Washington DC: U.S. Department of Health, Education and Welfare, 1980.
5. National Center for Health Statistics (NCHS). Long Term Health Care. Minimum Data Set. Report of the National Committee on Vital and Health Statistics. Washington DC: U.S. Department of Health, Education and Welfare, 1980.
6. United Nations. Statistical Office. The Role of the NHSCP in Providing Health Information in Developing Countries. NHSCP Technical Study No. 3. New York: United Nations, 1983.
7. Carlson B. Using surveys for management of health. Presentation given at the Xth Scientific Meeting of the International Epidemiological Association Meeting August 1984, Vancouver, Canada.
8. Lynge E. Regional trends in incidence of cervical cancer in Denmark in relation to local smear-taking activity. International Journal of Epidemiology, Vol. 12, No. 4, p. 405-413, 1983.
9. Mühling P., Bory J. und Haupt H. Einfluss der luftbelastung auf Atemwegserkrankungen: Untersuchungen bei Säuglingen und Kleinkindern. Staub. Reinhaltung der Luft, Band 45, Nr. 1, 1985.
10. Kerek-Bodden E., Schach E., Schach S., Schwartz F.-W., Wagner P. and Robra B.P. Ambulatory medical care and its role in the health care system. In: v. Eimeren W., Engelbrecht R. and Flagle Ch., Eds. Third International Conference on Systems Science in Health Care. Heidelberg : Springer, 1984.