International Journal of Epidemiology
© International Epidemiological Association 1996

Vol. 25, No. 2
Printed in Great Britain

A Framework for Evaluation of Secondary Data Sources for Epidemiological Research

HENRIK TOFT SØRENSEN,* SVEND SABROE** AND JØRN OLSEN†

Sørensen H T (Department of Internal Medicine V, Aarhus University Hospital DK-8000 Aarhus C, Denmark) Sabroe S and Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology* 1996; **25:** 435–442.

Background. As part of the development in information technology, increasing amounts of health care data are available for epidemiological research.

Methods. In this review, we discuss the following factors affecting the value of secondary data in research: 1) completeness of registration of individuals, 2) the accuracy and degree of completeness of the registered data, 3) the size of the data source, 4) the registration period, 5) data accessibility, availability and cost, 6) data format, and 7) possibilities of linkage with other data sources (record linkage).

Results and Conclusion. The importance of these issues depends on the use of the data and on the problems they have to address. If the evaluation is satisfactory with respect to the above-mentioned factors relevant to the particular study, the data source could be a very cost-effective way of solving the research problem.

Keywords: evaluation, validity, accuracy, information systems, data sources, records

Development in technology has led to a considerable increase in the number of individual-based data sources, registers, data bases, and information systems that may be of value in epidemiological research, and the number of studies that are based on secondary data may be expected to increase. Secondary data in research are data which have not been collected with a specific research purpose.[1] Such data are often collected for: 1) management, claims, administration and planning;[2] 2) evaluation of activities within health care;[2] 3) control functions;[3,4] and 4) surveillance or research.[5,6]

The main advantage of using secondary data sources is that they already exist; the time spent on the study is therefore likely to be considerably less than the time spent on studies that use primary data collection. Furthermore, the costs of the project are reduced markedly, as is the waste of data, compared with collection of primary data. Other advantages include the size of the sample, its representativeness, and the reduced likelihood of bias due to, for example, recall, non-response and effect on the diagnostic process of attention caused by the research question.[1,7]

The disadvantages of secondary data are related to the fact that their selection and quality, and the methods of their collection, are not under the control of the researcher, and that they are sometimes impossible to validate.

Despite comprehensive use of secondary data sources, the literature concerning this is relatively modest.[7–24] The purpose of this article is to address issues that are of importance to the use of a secondary data source for research, and to illustrate the subject matter with some examples.

Individual-based data sources usually consist of a series of records for each individual, each record containing several items of information, but since data collection and compilation have usually not been done with a current research purpose in mind, secondary data will usually not cover all aspects of interest.

As in all research, planning a study should aim at reducing both systematic and random errors. Any study based on secondary data should be designed with the same rigour as other studies, i.e. specifying hypotheses

* Department of Internal Medicine V, Aarhus University Hospital, DK-8000 Aarhus C, Denmark.
** Department of Epidemiology and Social Medicine, University of Aarhus, DK-8000 Aarhus C, Denmark.
† The Danish Epidemiology Science Centre, The Steno Institute of Public Health, University of Aarhus, DK-8000 Aarhus C, Denmark.
Reprint requests to: Henrik Toft Sørensen, The Danish Epidemiology Science Centre, Nørrebrogade 44, Building 2c, Aarhus University Hospital, DK-8000 Aarhus C, Denmark.

TABLE 1 *Factors affecting the value of secondary data in epidemiological research*

1. Completeness of registration of individuals
   a. Comparing the data source with one or more independent reference sources
   b. Comprehensive records review
   c. Aggregated methods
2. The accuracy and degree of completeness of variables
   a. Precision
   b. Validity
3. The size of the data sources
4. Registration period
5. Data accessibility, availability and cost
6. Data format
7. Record linkage

and estimating sample size to get valid answers. The available literature on the value of secondary data sources has mainly focused on completeness of registration of individuals and on accuracy of the variable registrations.[8,9] How good secondary data need to be depends of course on the research question in mind.

USE OF SECONDARY DATA IN RELATION TO TYPE OF STUDY

A. If secondary data are used for comparing occurrence data over time or between different populations, the outcome data should be complete. If they are not, the validity of diagnosis, as measured by the sensitivity and specificity, must be the same over time or in the populations to be compared. How close these validity measures have to be depends on the level of the desired detection and on the changes in the true underlying disease parameters. Even a very incomplete registration of infectious diseases will be able to demonstrate major epidemics of common diseases such as influenza and measles. The same is not necessarily true for congenital malformations for which changes in estimated occurrence data often reflect diagnostic differences rather than true changes in the disease prevalence.

B. If the aim is not to describe absolute occurrence data but relative rates in different populations, such a comparison may provide effect measures in the right direction despite incomplete registration as long as the putative cause has no impact on the diagnostic process. Non-differential misclassification will usually bias relative measures, such as the relative risk and odds ratio, towards the value of one; particularly in the case of low specificity for rare diseases. A low sensitivity, but of similar magnitude, will bias effect measures based on absolute values obtained by subtracting occurrence data in one population from occurrence data in another. In order to obtain a high specificity, all those with the diagnosis in question may only be accepted after close scrutiny, e.g. by going through existing medical files, asking for additional information, etc. The principle is similar to applying screening tests in sequence when only test positives are screened twice. Should it be possible to exclude everyone without the diagnosis, the case ascertainment will have a specificity of 1, making all relative measures unbiased, such as relative risks or relative prevalence ratios (though not odds ratios and not effect measures based on differences). The only price to pay for such a strategy is a reduction of the power of the study.

C. The strategy will of course not work if the aim is to estimate the frequency of disease occurrence. Sequential testing of test positives only will eventually lead to a prevalence of zero if each test has a sensitivity less than one. Sequential testing of only test negatives will in the same way lead to a prevalence of one if the tests have a specificity less than one. If validation aims at estimating the prevalence of the disease in the population of interest, the sensitivity and specificity of the recording need to be estimated. Let the estimated prevalence (based on inaccurate recording) be EP, the true prevalence P, then

$$EP = P \times sens + (1 - p) \times (1 - spec) \text{ and}$$
$$P = (EP + spec - 1)/(sens + spec - 1)$$

The formula illustrates that the frequently used strategy of checking only positive recordings will reduce the prevalence. Should all non-cases be removed in the check up, then the formula is ($P = EP/sens$ and $EP = P \times sens$).

A proper strategy has to include recorded cases as well as non-cases, and the sampling fractions depend on cost considerations.[25] Formulas for the predictive value of a positive test and the true prevalence of the disease when the tests are used in sequence are found in Olsen.[26]

D. If the aim is to describe the prognosis for a given disease, in order to apply it in general, the diagnosing must be unrelated to the prognosis; this is rarely the case. The most severe cases are usually under closer scrutiny, and Berkson's bias[27] is more likely to draw attention to patients suffering from several diseases. Unbiased estimates are difficult to reach and many outcome measures will be confounded by the indication for treatment and hospitalization.

FACTORS AFFECTING THE VALUE OF SECONDARY DATA

Completeness of Registration of Individuals

By this we mean the proportion of individuals in the target population which is correctly classified in the data source. In this respect, it is important to know whether the data source is population-based, like many of the European health insurance systems,[2] or whether it has been through one or more selection procedures (for instance in Medicaid which may exclude large groups of people who were included in previous years).[22,23] Furthermore, it is essential to know whether the denominator population is stable.

Methods for evaluation of completeness can be divided into three groups.

A. An estimate of the degree of completeness can be obtained by comparing the data source with one or more independent reference sources, in which the whole (total) or part (partial) of the target population is registered.[5,8] The comparison is made case by case.

Example: In a study on meningococcal disease in general practice, several patients with meningococcal meningitis were identified from the official Danish surveillance systems.[5] The identified cases were compared with information from hospital records, including information about culture of blood and spinal fluid from the departments of bacteriology, and it was discovered that, of 180 cases notified to the Department of Public Health during the study period, only 170 complied with the criteria for meningococcal disease. A further seven un-notified people with meningococcal disease were included from the records of the regional Microbiological Department, which was used as a reference source.

In this study the evaluation was made manually, which is very time-consuming, but computerized individual comparisons have been carried out in other studies.[24]

Cases escape even the best registration systems. Some researchers have therefore compared different independent data sources, and the missing cases have been estimated in a capture-recapture model.[28,29]

The degree-of-completeness concept is closely linked to the concept of sensitivity. The validity concept of the registration of cases as applied in the literature in the field of validation of registers (see next section); the ratio between the number of correctly registered individuals (e.g. meningococcal disease) and all those registered (e.g. everyone registered with meningococcal disease) is, however, closely linked to the concept of the predictive value of a positive registration. A comparison between two data sources alone does not provide the opportunity to estimate specificity. It can be assumed though that specificity will be close to one if the background population is big and the disease rather rare. A capture-recapture analysis offers the opportunity to estimate the number of cases not registered in the data sources, and if the background population is known as well, an estimate of the specificity can be obtained (Table 2).

TABLE 2 *Terminology in relation to evaluation of data source 1. Data source 2 is used for comparison, but is seldom a gold standard and cannot immediately be used as a reference for the true disease occurrence in a population. A prerequisite in the design of the table is to go through records ensuring that the cases fulfil the criteria of being cases*

Data source 1	Data source 2		
	Registered cases	Non-registered cases	
Registered cases	a	b	a + b
Non-registered cases	c	d	c + d
	a + c	b + d	

The background-population whereupon cases arise is called BP.
d can be estimated in a capture-recapture analysis (ref. 29). Thus the estimated total number of cases is a + b + c + d in BP.

Evaluation of data source 1:
The degree of completeness in data source 1 is usually calculated as = a/(a + c)
Estimate of register-based sensitivity in data source 1 = (a + b)/(a + b + c + d)
Estimate of register-based specificity in data source 1 = (BP − (c + d))/BP

B. Comprehensive records' review methods are used particularly in hospital discharge systems. All the cases should be registered, but some will probably be misclassified because of inaccurate or incomplete coding of diagnoses.

Example: In a Danish study on the occurrence and causes of anaphylactic shock outside hospital,[30,31] results were based on a population-based hospital discharge register. All records with a discharge diagnosis of anaphylactic shock were studied. In addition, records with a diagnosis of allergic and toxic reactions, adverse reactions to drugs, and shock not caused by cardiovascular disease or trauma were reviewed. Twenty records were identified. The recorded diagnoses at discharge were as follows: anaphylactic shock (n = 12), other, mainly bee sting (n = 8). Searching the anaphylactic shock diagnosis alone would have given a completeness rate of only 60%.

Several types of problems limit the usefulness of discharge diagnoses:[32] a) variations in coding, b) errors in coding, c) incompleteness in coding, especially of

comorbidities, d) limits in the specificity of available codes, e) errors and variation in clinical diagnosis. It is important to underline that discharge systems are often event-based and not person-based.

C. In aggregated methods the total number of cases in the data source is compared with the total number in other sources, or the expected number of cases is calculated by applying epidemiological rates from demographically similar populations or by simulation. Simulation uses the information system to simulate patterns of incomplete reporting to examine the possible effect on a specific dependent variable.[8,33]

Example: The aim of a Danish study was to report incidence of liver diseases based on the population-based nationwide Hospital Register.[34] The study included 512 cases of toxic hepatitis during the 5-year period from 1981 to 1985. During the decade 1978–1987 the Danish Board of Adverse Reactions to Drugs received 1100 reports on hepatotoxicity, and it was concluded that the figures for the periods were in close agreement.

The demands for completeness and representativeness depend on the research question. For several analytical studies the degree of completeness may be less important than whether the misclassification is random or differential. Since valid measures of effect size only depend on the odds of exposed to non-exposed among cases and controls, not the completeness of the case ascertainment, incomplete case ascertainment may be critical in a follow-up study, but less problematic in a case-control design. As long as the case identification is unrelated to the exposure of interest, a case registry may be used as a valid source of candidates for a case-control study.

Usually no reference standard for the evaluation of secondary data sources exists; thus the degree of completeness will often be given as the degree of agreement with one or more reference data sources.[8] Selection for participation in a study is often based on a certain diagnosis, or on patients treated with a special drug, and it is in relation to the selection criteria that the degree of completeness is evaluated.

It is also useful to pose two questions concerning the degree of completeness: a) does the information system cover all who are diagnosed with the disease in question (all who are eligible for the information system)? and b), more ambitiously, does the information system cover all the target populations with the diagnoses in question?

The Accuracy and Degree of Completeness of the Registered Data

The individual record will often contain several variables apart from the one by which the person is identified; for example, the results of certain diagnostic tests, diagnoses, age, gender, contacts with medical doctors, and demographic data.

Errors of variables can be divided into two groups: 1) random errors and 2) systematic errors. To secure the accuracy of a study it is necessary to try to reduce the occurrence of both types.

The concept of precision is complementary to the concept of random error. Validity is the extent to which the study measures what it is intended to measure. Lack of validity is referred to as bias or systematic error, and validity in the context of assessing data quality in information systems may be defined as the rate of cases in the information system with a given characteristic, which truly have this attribute.[8] Consequently, validity of the registration of cases is number of cases fulfilling the criteria for being a case/number of registered cases, and it is then close to the concept of the predictive value of a positive registration.

The accuracy of such registered data will often have to be evaluated by comparison with independent external criteria.[8] With respect to completeness and accuracy, the diagnoses will often have to be compared with operational criteria by going through records.[5,35] However, other data are based on an evaluation, e.g. the result of an ECG or of x-ray examinations, as well as on certain diagnoses. Thus, there is not necessarily any reference standard, and it may be important to examine the reproducibility (inter and intra-observer variation).[36–39] Furthermore, the evaluation also includes the extent of missing data, since a significant degree of missing and incomplete data negates the value of the source.[11,40] For each single variable it should be considered whether missing information means that exposure or outcome has not taken place or whether the variable represents a missing value. Inaccurate or missing data tend to bias associations toward the null hypothesis rather than to cause spurious associations, as long as they occur in equal proportion in the groups to be compared.[41]

Example: The aim of a Danish study was to investigate the validity of information on abortion in the Danish Hospital Register.[42] The information in the register was compared with data in the discharge records (359 records from 31 hospitals). Agreement between the two data sources was 92–100% for administrative data (personal identification number, hospital identification, data of hospitalization and ICD codes). Disagreement was 31–54% between the diagnosis in Latin and the number code of the diagnosis in the discharge records.

The above-mentioned points are of special importance to the precision and validity, and possibly the degree of misclassification of data in an investigation.

Another recurring problem with data base studies concerns missing information about potential confounders;[43,44] a potential confounder which is often lacking in secondary data sources is smoking. It has been calculated that even when one is dealing with large effects (i.e. risk ratios of ≥2) different smoking habits will usually only be able to 'explain' part of the association.[44]

In summary, data quality problems can be categorized as follows: 1) errors in the data set may reflect incorrect data entry or lack of entry of available information, and 2) the original source of information may be correctly entered into the data source but may not reflect the true condition or characteristic of the subject.

The Size of the Data Source

It is essential to know how many people and how many variables are registered in the data source. Furthermore, it may be relevant to know the distribution of the various variables since this may be of importance in designing the study to provide it with proper dimension.[45] Use of restriction and matching in the control of confounding factors and sources of selection often require progressively more subjects as the number of matching variables increases.

If the data source is very large, even small associations will give statistically significant results. It is therefore essential to relate the size of the data source to the clinical relevance of any difference, rather than to look at the *P*-values. In registry studies, as in other types of study, it is more important to calculate the confidence interval around the estimated difference than to calculate the *P*-values.[11,46]

Registration Period

Often data sources only contain cross-sectional registrations,[47] which reduce the possibility for analytical studies. With respect to longitudinal studies, information concerning the registration period(s) is essential for the design in order to relate exposure and effect to possible induction and latent periods. The induction period is the period required for a specific cause to produce disease, the latent period is the delay between the exposure and the period of manifestation of the disease. Data sources with observation periods of a few years will seldom be suitable for aetiological cancer research. The length of registration may also be important in ascertaining cases where the diagnosis is delayed, e.g. congenital heart diseases are often not diagnosed until after the neonatal period.

Furthermore, codes, and even the layout of records, are often changed periodically. Changes in diagnostic criteria and classifications (e.g. the recent change from ICD-9-CM to ICD-10 disease classification system) frequently cause problems when comparing data over longer periods.

Example: Based on data from the Danish Cancer Register in the period 1943–1985 it has been shown empirically that the increase in the incidence of registered primary liver cancer in Denmark may be ascribed to changes that have taken place in diagnostic procedures, which resulted in changes in the classification of liver cancer from unspecified liver cancer to primary liver cancer.[48]

This type of problem is obviously of importance in studying time trends and when comparing groups for which the observation periods do not run in parallel.

Data Accessibility, Availability and Cost

It is often not clear who owns the data and who has the right to use them (accessibility).[49] It is important to clarify these points and to find out which authorities should approve the use of the data for research purposes. It is well-known that general practitioners and hospitals do not always respond or do not accept the use of their records for research.[3,37,50] Records may have been destroyed.[51]

It is also important to know the financial costs of using the data and for having them made available. Data are sometimes available on-line, but more often special programmes are needed. Information on data confidentiality is also essential in order to ensure protection of confidentiality of data on individuals which are reported to the data sources, so that information on those registered cannot reach unauthorized third parties.[52] Confidentiality can often be maintained by using multiple passwords for data access and by using abbreviated identifiers in the researchers' data.[13]

Data Format

Data will often be in the form of paper records from hospitals[5] or be computerized, e.g. many health insurance data.[2,11,18,21–23] Even computerized records can be formatted or structured in such a way that their use is made difficult for research, e.g. inappropriate format of variables (e.g. diagnostic categories, age bands).

Possibilities of Linkage with Other Data Sources (Record Linkage)

Important research results have been obtained by linkage of different data sources.[53,54] Record linkage techniques can help to identify the same person in different files. There are for example several computerized health care data bases in North America.[10,18,21–23,35,55,56]

By using computerized billing records, drug exposure was linked to files which included diagnosis (internal record linkage). Linkage is done via a personal identification number. Consequently, these data bases constitute powerful tools for drug evaluation. However, a complete high-quality record linkage may not always be possible. The best population-based data sources are probably the extensive data linkage networks in Scandinavia, where each person is assigned a unique personal registration number at birth (CPR-number), allowing record linkage between several independent data systems and vital statistical registers (external record linkage).[57–59]

If, as is usual, there is no personal identification number, the linkage must be done on other types of joint identification information, such as name, date of birth, diagnosis, etc. The identifiers should be unique, permanent, universal and available. The problem is to overcome errors in the identifying information and to fully exploit use of the discrimination power of various items that identify the person. Variation in spelling of family names often occurs (4–5%). One way of dealing with this is the Russel Soundex Code, in which surnames are reduced to the first letter followed by numeric digits.[60]

Example: In Denmark information on all cancers that have been diagnosed since 1943 is reported to a central register (the Danish Cancer Register) and this registry is updated by annual linkage to death certificates for inclusion of unreported cancer cases. In a study on patients with cancer of the cervix at three radiotherapy centres in Denmark, 5674 cases were diagnosed in the period after 1943.[61]

Linkage of the records in the Cancer Register with the complementary cervical cancer file was carried out using month and year of birth, surname, first forename and year of diagnosis. Matching was done by both computerized and visual means. For the period of diagnosis 1943–1966, 2.2% of cases were not identified in the Cancer Registry file. Hospital records were then scrutinized for unidentified patients. There were different errors: different date for diagnosis, errors in date of birth and name. After correcting for these errors the true underreporting at the central cancer register was 0.9%.

ETHICS AND DATA CONFIDENTIALITY

The European Union (EU) has made a proposal for a directive on the protection of people in connection with the handling of personal data information.[62] The proposal contains extensive protective precautions with respect to the registered citizens, including the demand for consent on a well-informed basis and a duty to inform the people concerned when personal data are used. The passing of this proposal will pose a serious problem to research based on information systems.[63–65] The proposal has just been examined by the EU, but the final text is not yet known.[66] Several of these systems are made for administrative purposes focusing on individuals, while researchers are interested in groups and their interaction. The proposal only affects research and has no protective effect on misuse by the administrative apparatus which collected the data.

CONCLUSIONS

It is a major advantage to be able to use existing data sources, with large amounts of information, which are relatively easily available for research purposes. Often millions of person-years of experience in the data bases will be available, which would be impossible to collect in prospective studies.

Data bases can be used as a sampling frame to select study populations, and to collect information on exposure, diseases, and sometimes confounders.[2]

Existing documentation should be critically reviewed to assess the appropriateness of the data for their intended use, and if such documentation does not exist, the researcher must evaluate the data source. This involves the protocols, record layout and codes, data entry instructions, published material, analyses, technical reports, and the carrying out of appropriate completeness and validity studies, all with respect to the specific context of the study.

If the evaluation is satisfactory with respect to each of the above-mentioned factors, the data source in question may be of value for solving the research problems, or at least, to provide a first evaluation of a given research problem which may set priorities for subsequent in-depth studies. Usually such studies can be done without any risk of disclosure and to the benefit of the Public Good.

ACKNOWLEDGEMENT

This research was supported by Fonden vedr. finansiering af forskning i almen praksis og sundhedsvæsenet iøvrigt (j.nr. 2–01–125).

REFERENCES

[1] Hearst N, Hulley S B. Using secondary data. In: Hulley S B, Cummings S R (eds). *Designing Clinical Research*. Baltimore: Williams & Wilkins, 1988, pp. 53–62.

[2] Sørensen H T. Re: A compendium of public health data sources. *Am J Epidemiol* 1992; **135**: 325–26.

[3] Sørensen H T, Rasmussen H H, Ejlersen E, Møller-Petersen J F, Hamburger H, Olesen F. Epidemiology of pain requiring strong analgesics outside hospital in a geographically defined population in Denmark. *Dan Med Bull* 1992; **39:** 464–67.

[4] Sørensen H T, Ejlersen E, Rasmussen H H, Møller-Petersen J F, Hamburger H, Olesen F. One year follow-up after the first prescription of strong analgesics outside hospital. *Int J Risk Safety Medicine* 1994; **4:** 209–13.

[5] Sørensen H T, Møller-Petersen J, Krarup H B, Petersen H, Hansen H, Hamburger H. Diagnostic problems with meningococcal disease in general practice. *J Clin Epidemiol* 1992; **45:** 1289–93.

[6] Sørensen H T, Møller-Petersen J, Krarup H B, Petersen H, Hansen H, Hamburger H. Early treatment of meningococcal disease. *Br Med J* 1992; **365:** 774.

[7] Roos L L, Roos N P, Fisher E S, Buholz T A. Strengths and weakness of health insurance data systems for assessing outcomes. In: Gelijns A C (ed.). *Medical Innovation at the Crossroads. Volume I. Modern Methods of Clinical Investigations.* Washington DC: National Academy Press, 1990, pp. 47–67.

[8] Goldberg J, Gelfand H M, Levy P S. Registry evaluation methods: a review and case study. *Epidemiol Rev* 1980; **2:** 210–20.

[9] Stone D H. A method for the validation of data in a register. *Public Health* 1986; **100:** 316–24.

[10] Roos L L, Mustard C A, Nicol J P *et al.* Registries and administrative data: organization and accuracy. *Med Care* 1993; **31:** 201–12.

[11] Connel F A, Diehr P, Hart L G. The use of large data bases in health care studies. *Ann Rev Public Health* 1987; **8:** 51–74.

[12] Roos L L, Roos N P, Cageorge S M, Nicol J P. How good are the data? Reliability of one health care data bank. *Med Care* 1982, **20:** 266–76.

[13] Roos L L, Nicol J P, Cageorge S M. Using administrative data for longitudinal research: comparisons with primary data collection. *J Chron Dis* 1987; **40:** 41–49.

[14] Armstrong B K, White E, Saracci R. *Principles in Exposure Measurement in Epidemiology.* Oxford: Oxford Medical Publications, 1992, pp 197–235.

[15] Pryor D B, Califf R M, Harrell F E *et al.* Clinical data bases. *Med Care* 1985; **23:** 623–47.

[16] Hierholzer W J. Health care data, the epidemiologist's sand: comments on the quantity and quality of data. *Am J Med* 1991; **91(Suppl. 3B):** 21–26.

[17] Feinleib M. Data bases, data banks and data dredging: the agony and the ecstasy. *J Chron Dis* 1984; **37:** 783–90.

[18] Fisher E S, Baron J A, Malenka D J, Barrett J, Bubolz T A. Overcoming potential pitfalls in the use of Medicare data for epidemiologic research. *Am J Public Health* 1990; **86:** 1487–90.

[19] Roos L L, Sharp S M, Cohn M M. Comparing clinical information with claims data: some similarities and differences. *J Clin Epidemiol* 1991; **44:** 881–88.

[20] Romano P S, Roos L L, Luft H S, Jollis J G, Doliszny K and The Ischemic Heart Disease Patient Outcomes Research Team. A comparison of administrative versus clinical data: coronary artery bypass surgery as an example. *J Clin Epidemiol* 1994; **47:** 249–60.

[21] Lauderdale D S, Furner S E, Miles T P, Goldberg J. Epidemiologic uses of Medicare data. *Epidemiol Rev* 1993; **15:** 319–27.

[22] Bright R A, Avorn J, Everitt D E. MEDICAID data as a resource for epidemiologic studies: strengths and limitations. *J Clin Epidemiol* 1989; **42:** 937–45.

[23] Ray W A, Griffin M R. Use of MEDICAID data for pharmacoepidemiology. *Am J Epidemiol* 1989; **129:** 837–49.

[24] Roos L L, Sharp S M, Wajda A. Assessing data quality: a computerized approach. *Soc Sci Med* 1989; **28:** 175–82.

[25] Irwig L, Glasziou P P, Berry G, Chock C, Mock P, Simpson J M. Efficient study designs to assess the accuracy of screening tests. *Am J Epidemiol* 1994; **140:** 759–60.

[26] Olsen J. Measurement techniques. In: Olesen J. *Headache Classification and Epidemiology. Frontiers in Headache Research.* New York: Raven Press, 1994.

[27] Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biomet Bull* 1946; **2:** 47–53.

[28] Hook E B, Regal R R. The value of capture-recapture methods even for apparent exhaustive surveys. *Am J Epidemiol* 1992; **135:** 1060–67.

[29] Bishop Y M M, Flenberg S E, Holland P W. Estimating the size of a closed population. In: *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press, 1975; pp. 229–56.

[30] Sørensen H T, Nielsen B, Nielsen J Ø. Anaphylactic shock occurring outside hospitals. *Allergy* 1989; **44:** 288–90.

[31] Sørensen H T, Nielsen B, Nielsen J Ø. Anaphylaxis. *N Engl J Med* 1991; **325:** 1658.

[32] Steinberg E P, Whittle J, Anderson G F. Impact of claims data research on clinical practice. *Int J Technol Assess Health Care* 1990; **6:** 282–87.

[33] Christensen K, Fogh-Andersen P. Isolated cleft palate in Danish multiple births, 1970–1990. *Cleft Palate Craniofac J* 1993, **30:** 469–74.

[34] Almdal T P, Sørensen T I A and The Danish Association for the Study of the Liver. Incidence of parenchymal liver diseases in Denmark, 1981 to 1985: analysis of hospitalization registry data. *Hepatology* 1991; **13:**650–55.

[35] Tennis P, Bombardier C, Malcolm E, Downey W. Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan hospital separations database. *J Clin Epidemiol* 1993; **46:** 675–83.

[36] Herrmann N, Cayten C G, Senior J, Staroscik R, Walsh S, Woll M. Interobserver and intraobserver reliability in the collection of emergency medical services data. *Health Serv Res* 1980; **15:** 127–43.

[37] Sundhedsstyrelsen. *Evaluering af landspatientregistret* 1990. Copenhagen: National Board of Health, 1993.

[38] Boyd N F, Pater J L, Ginsburg A D, Myers R E. Observer variation in the classification of information from medical records. *J Chron Dis* 1979; **32:** 327–32.

[39] Beard C M, Yunginger J W, Reed C E, O'Connell E J, Silverstein M D. Interobserver variability in medical record review: an epidemiological study of asthma. *J Clin Epidemiol* 1992; **45:** 1013–20.

[40] Iezzoni L I, Foley S M, Daley J, Hughes J, Fisher E S, Heeren T. Comorbidities, complications, and coding bias. *JAMA* 1992; **267:** 2197–203.

[41] Rothman K J. *Modern Epidemiology.* Boston: Little Brown and Company, 1986.

[42] Schmidt L, Damsgaard M T, Nielsen J M. Evaluering af landspatientregistret. *Ugeskr Læger* 1989; **151:** 3478–82.

[43] Brownson R C, Davis J R, Chang J C, DiLorenzo T M, Keefe T J, Bagby J J Jr. A study of the accuracy of cancer risk factor information reported to a central registry compared

to that obtained by interview. *Am J Epidemiol* 1989; **129:** 616–24.

[44] Axelson O. Aspects of confounding and effect modification in the assessment of occupational cancer risk. *J Toxicol Environ Health* 1980; **6:** 1127–31.

[45] Strom B L. Sample size considerations for pharmacoepidemiologic studies. In: Strom B L (ed). *Pharmacoepidemiology*. New York: Churchill Livingstone, 1989, pp. 27–37.

[46] Walker A M. Reporting results of epidemiologic studies. *Am J Public Health* 1986; **76:** 556–58.

[47] Gable C B. A compendium of public health data sources. *Am J Epidemiol* 1990; **131:** 381–94.

[48] Andersen I B, Sørensen T I A, Prener A. Increase in incidence of disease due to diagnostic drift: primary liver cancer in Denmark, 1943–85. *Br Med J* 1991; **302**: 437–40.

[49] Smith G D. Increasing the accessibility of data. *Br Med J* 1994; **308:** 1519–20.

[50] Tilley B C, Barnes A B, Bergstrahl E *et al.* A comparison of pregnancy history recall and medical records: implications for retrospective studies. *Am J Epidemiol* 1985; **121:** 269–81.

[51] Horwitz R I, Yu E C. Assessing the reliability of epidemiologic data obtained from medical records. *J Chron Dis* 1984; **37:** 825–31.

[52] Coleman M P, Muir C S, Ménégoz F. Confidentiality in the cancer registry. *Br J Cancer* 1992; **62:** 1138–49.

[53] Roos L L, Wajda A. Record linkages strategies. *Meth Inform Med* 1991; **30:** 117–23.

[54] Roos L L, Wajda A, Nicol J P. The art and science of record linkage: methods that work with few identifiers *Comput Biol Med* 1986; **16:** 45–57.

[55] Tennis P, Andrews E, Bombardier C *et al.* Record linkage to conduct an epidemiologic study on the association of rheumatoid arthritis and lymphoma in the province of Saskatchewan, Canada. *J Clin Epidemiol* 1993; **46:** 685–95.

[56] Rodriquez L A G, Walker A M, Gutthann S P. Nonsteroidal antiinflammatory drugs and gastrointestinal hospitalizations in Saskatchewan: a cohort study. *Epidemiology* 1992; **3:** 337–42.

[57] Kramer M S. *Clinical Epidemiology and Biostatistics*. Berlin: Springer-Verlag, 1988.

[58] Frisch M, Olsen J H, Bautz A, Melbye M. Benign anal lesions and the risk of anal cancer. *N Engl J Med* 1994; **331:** 300–02.

[59] Sørensen H T, Larsen B O. A population-based Danish data resource with possible high validity in pharmacoepidemiological research. *J Med System* 1994; **18:** 33–38.

[60] Newcombe H B, Kennedy J M, Axford S J, James A P. Automatic linkage of vital records. *Science* 1959; **130:** 954–59.

[61] Storm H H. Completeness of cancer registration in Denmark 1943–1966 and efficacy of record linkages procedures. *Int J Epidemiol* 1988; **17:** 44–49.

[62] Commission of the European Communities. *Amended proposal for a council directive on the protection of individuals with regard to processing of personal data on the free movement of such data.* Brussels: CEC, 1992. (COM (92) 422 final syn 287).

[63] Knox E G. Confidential medical records and epidemiological research. *Br Med J* 1992; **304:** 727–28.

[64] Protecting individuals: preserving data (editorial) *Lancet* 1992; **339:** 784.

[65] Lynge E. European directive on confidential data: threat to epidemiology. *Br Med J* 1994; **308:** 490.

[66] Lynge E. New draft on European directive on confidential data. *Br Med J* 1995; **310:** 1024.

(Revised version received July 1995)