Eur J Epidemiol
DOI 10.1007/s10654-013-9873-0

METHODS

When the entire population is the sample: strengths and limitations in register-based epidemiology

Lau Caspar Thygesen · Annette Kjær Ersbøll

Received: 20 August 2013 / Accepted: 12 December 2013

Abstract Studies based on databases, medical records and registers are used extensively today in epidemiological research. Despite the increasing use, no developed methodological literature on use and evaluation of population-based registers is available, even though data collection in register-based studies differs from researcher-collected data, all persons in a population are available and traditional statistical analyses focusing on sampling error as the main source of uncertainty may not be relevant. We present the main strengths and limitations of register-based studies, biases especially important in register-based studies and methods for evaluating completeness and validity of registers. The main strengths are that data already exist and valuable time has passed, complete study populations minimizing selection bias and independently collected data. Main limitations are that necessary information may be unavailable, data collection is not done by the researcher, confounder information is lacking, missing information on data quality, truncation at start of follow-up making it difficult to differentiate between prevalent and incident cases and the risk of data dredging. We conclude that epidemiological studies with inclusion of all persons in a population followed for decades available relatively fast are important data sources for modern epidemiology, but it is important to acknowledge the data limitations.

Keywords Registers · Database management systems · Epidemiology · Bias · Nordic countries

L. C. Thygesen (✉) · A. K. Ersbøll
National Institute of Public Health, University of Southern Denmark, Øster Farimagsgade 5 A, 1353 Copenhagen K, Denmark
e-mail: lct@niph.dk

Research based on databases, medical records and registers are used more extensively than ever. During the eighteenth and nineteenth centuries few disease registers existed, with the Leprosy Registry in Norway [1] as the oldest followed by several tuberculosis registers [2]. In many countries, e.g. United Kingdom, Canada and Australia, research databases have been implemented [3]. The breakthrough in use of Nordic registers happened with the introduction of the unique personal identification number in 1964–1969 making individual-level linkage between registers possible in a reliable manner [4–7]. Administrative and research registers [8–10] were established with information about diseases, contact to the health care system, education and income.

An increasing amount of health care data is available for epidemiological research [8]. The importance of registration systems is not only due to the large amount of social and health events recorded, but even more to long follow-up time with data available for many years [4].

Numerous examples of novel results exist due to the possibility of performing register-based studies. One example is case–control studies suggesting that induced abortions increased breast cancer risk [11, 12]. However, a Danish register-based cohort study including all Danish women born from April 1935 through March 1978 found no overall association between induced abortion and risk of breast cancer [13] based on data in the Register of Legally Induced Abortions [14] and the Danish Cancer Registry [15]. This study suggests that bias (e.g. recall or selection) might have influenced the results from the case–control studies.

Studies of detrimental effects of drugs taken by children or pregnant women on the unborn child could not be conducted in randomized trials due to ethical reasons. Furthermore, self-reported exposure data on prescribed

Published online: 10 January 2014

drugs may be influenced by recall bias. Register-based information on drug use, outcomes and co-morbidities offer important possibilities. An example is a register-based case–control study where researchers reported that oral contraceptive use in early pregnancy do not increase the risk of hypospadias in male offsprings [16].

Despite the increasing use of registers, there is no developed methodological literature on how to use and evaluate population-based registers and administrative databases.

One could argue that the only difference is that register-based epidemiological studies are based on another data source compared to traditional epidemiological studies based on surveys or clinical information. In some aspects this is correct: Any study based on data from registers should be designed with the same critical approach as studies based on other data sources, e.g. specifying hypotheses, estimating sample size, considering study design, and evaluating bias and statistical precision to obtain valid answers [9, 10]. The concepts of originality and credibility are essential in register-based epidemiology as in all medical research.

On the other hand, there are some noticeable differences between register-based epidemiological studies including the whole population and other epidemiological studies based on sampling within a population often including self-reported information on exposure, confounders and outcome. One difference is the connection between the research question and data [17, 18]. In traditional epidemiological studies the researcher collects her own data, while in register-based studies data are extracted from registers, where e.g. administrators have collected and entered data in the register. This limits the control of the data collection both with respect to variables collected, but also the content of variables, which is often collected with other aims than research. Dans [19] argues that data sets completed for billing purposes and constructed mainly by financial experts differ substantially from those constructed by clinicians caring for patients. Hsia et al. [20] have shown a diagnostic drift towards diagnoses with higher costs, which influences the validity of disease classifications in hospital systems. However, data have been collected prospectively in that information on exposure is not influenced by later diagnosis of disease, thereby minimizing the risk of recall bias [21]. Misclassification of exposure and outcome will often be non-differential because it will probably be the same for all population groups and will therefore tend to underestimate the true association.

Another difference is that registers are population-based with the possibility of including all persons in a defined population. This limits the influence of (and to some extent exclude) selection bias in register-based studies compared to studies based on a sample from a population, which furthermore could be influenced by sub-optimal participation rates.

Finally, traditional mathematically oriented statistical analyses with focus on sampling error as the main source of uncertainty may not be as relevant in register-based studies, because sampling error may not be the most appropriate when including the total population. Further, traditional statistical significance may not be of interest in massive number of observations in the data sets, because even practically unimportant differences easily become statistically significant within large data sets.

In this paper we define relevant concepts, present strengths and limitations of register-based epidemiological studies and discuss epidemiological biases especially to be considered in register-based studies.

Definitions

In this article we use the definition of a register given by the United Nations Economic Commission for Europe and Wallgren and Wallgren in that a register should be a complete listing and each individual should be identifiable for updating [22, 23]. Samples and anonymized complete listings of individuals are therefore not registers. Base registers is one group of registers of great importance to statistical systems as they keep stock of the population at any given time and contain link to other base registers [5, 22, 23]. In all Nordic countries, at least three base registers are defined: register on persons (population register, linkage key is the personal identification number), business register (linkage key is the business identification number) and register on properties (real estate, buildings and dwellings, linkage key is the building and housing identification number) [5, 22, 23]. It is possible to link persons with firms and properties by these unique linking keys.

The title of this paper is the contradiction between a sample and the entire population. The sample of a register-based study has the possibility of encompassing the whole population of a given country in a given time period. Even though the sample of a register-based study includes all residents of a country (e.g. all Danes in a given period) this can be considered as a sample of a larger potential population over time and geography and thereby as a realization of a stochastic process. In medical research we often wish to generalize the results from one register-based study to other populations or the same population including future cases. The study population is the residents of a given population but the hypothetical target population is larger. The (finite) population at a given time might be considered as a sample (i.e. a realization) of a larger theoretical population sometimes referred to as the super-population [24].

Traditionally, statistical inference (e.g. hypothesis testing and *p* values) is defined as drawing conclusions about a population based on a sample from the population with the error term based on sampling error. In register-based studies we still consider the outcome measured with error and predict the outcome or future cases using an underlying probability model. The unobserved (latent) variables and the uncertainty of the outcome introduce the error terms. When including data from a number of years in register-based studies the error term might be based on year-to-year variations by introducing year as a random term in the model. Finally, resampling-based tests can be used (e.g. randomization and permutation tests, bootstrapping and jackknifing) where data at hand are used and inference is based on repeated random allocation of the actual data values [25].

In conclusion, statistical inference is appropriate in register-based studies. However, the influence of chance is low because of the massive number of observations in the data sets.

Strengths of register-based epidemiology

Register-based epidemiological studies have several strengths (Textbox 1). Using registers in epidemiological studies can be seen as research economy in a broad sense in that if registers were not available the same studies could have been done, but with much higher costs [6, 21, 26]. This argument is true when registers already have been established for other reasons than research, but in situations where no registers are available, the initiation, development and maintenance of registers may have higher costs than data collection for one specific research study.

The first strength is that data already exist, which makes data collection faster and less expensive to conduct [8–10, 21, 26]. Furthermore, register-based studies often have large sample size and therefore great statistical power, which makes studies of rare exposures and outcomes possible [4, 6, 27]. A third strength is that registers typically are complete as far as the persons in the target population are concerned [9, 10, 26], which ensures representativeness and studies of associations in a real-world setup. The completeness minimizes the effects of selection bias due to non-response and loss to follow-up (attrition bias). This also makes it possible to focus on small sub-populations, e.g. persons in a specific area or with a special combination of socio-economic attributes. This is a great strength compared to surveys or health examinations, which often are influenced by sub-optimal participation rates [6].

Textbox 1 Strenghts of register-based data for research

1. Data already exist
2. Large sample size
3. Data are complete as far as the persons in the target population

- Limited/no selection bias
- No attrition bias
- Possibility to study rare exposure and outcome measures
- Information of exposures and outcomes for the whole population

4. Data are collected independently of research questions

- Prospective data collection
- No differential misclassification

5. Valuable time has passed

- Possibility to study diseases in families (generation studies)

6. Adjustment for confounders available to the whole population
7. Sometimes registers have the information of exposure and outcome of interest

A fourth strength is that the collection of data has been done independently of the study, and this often reduces various types of bias such as recall and influence of the diagnostic process determined by the study [9, 10, 26] assuring non-differential, independent classification errors. This also comes from the fact that the information is collected before the project ensuring prospective data collection.

Fifthly, valuable time has passed; many health problems manifest themselves many years after exposure and existing registers are thus especially valuable when studying diseases with a long latency period between exposure and disease manifestation and when inferences on induction and latent periods are needed [9, 10, 26]. This strength is vital for many health outcomes studied in modern epidemiology [6] and also makes studies of long-term trends in disease incidence possible. A specific aspect of this is diseases that occur in families, where registers make inter-generation studies and studies among siblings, half-siblings and twins possible [28].

A sixth strengths is that it is possible to adjust for some confounders available for the whole population [27] and that information on administrative conditions is often registered with very high completeness and validity [6, 29]. These variables include educational level, income, housing, family membership, hospitalizations, visits to general practitioners, reimbursement of drugs and vital status [4, 30]. Such data will often have higher validity than self-reported data.

Limitations of register-based epidemiology

Register-based epidemiological studies also have limitations, which are important to recognize (Textbox 2). Data in registers are pre-collected whereby necessary information may be unavailable, un-acquired, inaccurate or misclassified [27]. Data selection and quality are defined by the register and not controlled by the researcher [9, 10].

Textbox 2 Limitations of register-based data for research

1. Data are pre-collected by others than researchers
- Necessary information may be unavailable or misclassified
- Often hard to know exactly how data are generated
- Limited to use variables in register
- Variation in coding between persons and institutions
- Coding used in registers may not be detailed
2. Lack of confounder information
3. Missing data difficult to handle
- Difficult to know what missingness means
- Under-coverage
4. Low or unknown data quality
5. Left truncation
6. Data dredging and misleading post hoc analysis
7. Unimportant differences become statistical significant

The registers are "the administrators' view of the world" and the researcher is limited to use definitions from administrative practices [27]. Registers contain information on the citizens in relation to public administrators and researchers are distant from the actual data collection. Furthermore, research topics need to suit the registers. Finally, it is often hard to know exactly how data are generated.

When using data from registers for research purposes the researcher is limited to use the variables recorded and included in the registers. Furthermore, the researcher is limited to the level of detail and coding used in the registers. Sometimes coded diagnoses are not the most relevant and there may be variation in coding practice between persons, between departments, between institutions or over time [31], e.g. when using new coding systems or incomplete coding seen among seriously ill patients [32, 33].

Register data sometimes also lack important information, such as information not reported to the register or not registered due to low registration frequency. An illness can progress and diagnosing the disease might be done at a later stage. As an example the appropriateness of using the psychiatric diagnosis at the first psychiatric admission has been challenged [27]. Sudden declines in income are unrecorded. Changes in threshold for admission (e.g. deinstitutionalization) can change hospitalization rates, which could influence register-based incidence and prevalence data [27]. Register data only include information on legitimately earned money and approved educations. American studies have shown that there may be a diagnostic drift towards diagnoses that have higher costs [20].

Another limitation is lack of confounder information. Often registers only contain limited and unspecific confounder information. This limitation together with the fact that register-based studies often have great statistical power to detect small effect sizes makes register-based studies prone to confounding [34]. Sometimes information on potentially important confounding variables is available in registers only at a point in time that is not relevant to the question at hand [35]. Important examples in pharmacoepidemiology are healthy drug-use effect and healthy drug-adherer effect [34] and in clinical epidemiology information on clinical status is important. By creative and clinical use of ICD-codes one can get an impression of patient's clinical status [9, 10]. Several methods have been proposed to study the influence of confounding, e.g. sensitivity analyses based on an array of informed assumptions or external adjustment with confounder information from an internal or external sample, have been proposed [36], but only few studies have compared these methods [37]. Another method to control for measured and unmeasured confounders is use of instrumental variables [38, 39]. An instrumental variable is defined by two criteria: It causes variation in the treatment variable (exposure) and it does not have a direct effect on the outcome variable, only indirectly through the exposure. When this is satisfied it facilitates causal inference from observational studies even if some confounders are unmeasured [38, 39]. Instrumental variable methods are not suited for small sample sizes, which is usually not a matter of concern in register-based studies and the methods are an attractive possibility in register-based studies. An example is the register-based study of chemotherapy for advanced lung cancer among elderly in the Survival, Epidemiology, and End Result tumor registry, where regional variation in chemotherapy was used as an instrument [40]. The study showed that chemotherapy was associated with significantly lower mortality in this patient group and the study supported that the instrumental variable in fact was randomly distributed with regard to measured confounders.

A third limitation is missing data. For many variables in registers only few missing data are observed, but often it is not clear what missingness means. It could mean that an event (i.e. exposure or outcome) did not happen or could mean the same as a specific category, e.g. in the Income register a missing value means the lowest income. One important concept is under-coverage, e.g. among immigrants the highest educational level is often missing in Danish registers or educational attainment is often missing for persons with education taken abroad [29].

Documentation of registers and classifications used (metadata) are crucial for register-based studies. Metadata describe data and variables by giving definitions of populations, objects, variables, methodology and quality [23]. There is an increasing need for metadata in register-based research. In surveys and other datasets with own data collection this information is available, which is a great difference between register-based data and research project data.

A fourth limitation is evaluation of data quality. Often it is only possible to validate register data by themselves, e.g. through cross-tabulations between different registers [27], and it is often not possible to validate register data against a golden standard that cannot be established [9, 10].

A fifth limitation is that data from registers are truncated by start of registration (left truncation) [27]. As an example consider a register-based study of breast cancer incidence. A woman hospitalized for breast cancer before cancer registration was initiated will be classified as healthy. Any re-admission for breast cancer will be wrongly classified as an incident breast cancer case. This phenomenon will overestimate the incidence especially in the first years of registration, while the prevalence of disease will be underestimated especially for diseases with low morbidity and few contacts to hospitals.

The large number of available data may lead to data dredging and misleading post hoc analysis [9, 10]. The temptation to use registers for research because they are large and include a defined population is understandable [19, 28], but the process of first identifying the data, and then proceed to the question is not good science. However, explorative studies may use large databases for generating hypothesis for further studies.

Finally, unimportant differences may become statistical significant in large-scale register-based studies. Therefore, it is important to interpret not only the significance level but also the size of the risk estimates and evaluate whether they have public health or clinical relevance.

Epidemiological bias especially to consider in register-based studies

One important bias to consider especially in register-based studies is confounding due to lack of data of important confounders or only crude information on confounders (residual confounding) [28]. In the Nordic countries, information on socio-economic factors as educational level, income, occupation, transfer payments or housing conditions are available for all citizens. Studies have shown that socio-economic factors correlate with lifestyle factors such as smoking, physical activity and diet [41, 42] and that adjusting for the socio-economic factors may be proxy variables of lifestyle factors.

In some instances comorbidity is important to include in the analyses to adjust for differences in morbidity in order to make appropriate comparisons. This is especially important in studies of survival after diagnosis or effectiveness of treatment. Several indexes have been proposed [43], e.g. the Charlson's index [44] and its many different refinements [45, 46] or the Chronic Disease Score reflecting prescribed medications [47] also extended to administrative registers [43]. In general, it is important to have insights into the coding of secondary diagnoses in a hospital administrative system, e.g. in some countries the use of secondary diagnoses influence the payment for a given treatment, which may result in over-reporting of some unimportant diseases thereby overestimating the comorbidity of a given patient. It is also important to consider which comorbidities could influence the association studied and maybe include just indicators of important comorbidities. Often the indexes are not specific enough for the particular study and it has been suggested that weights for a given comorbidity index should be derived by study-specific weights [43].

Sensitivity analyses can be used to evaluate the influence of unmeasured confounders [48] and is suggested being used to evaluate the robustness of the results with respect to unmeasured confounding [49]. Schneeweiss [36] suggests using external information or sensitivity analyses in database studies of therapeutics to evaluate the influence of unmeasured confounders. He shows that the inclusion of external information on confounders could be implemented in a case study of COX2 inhibitors and myocardial infarction showing that five confounders only slightly changed the unadjusted result [37].

Often information on exposure is only a proxy for the variables of interest probably reflecting misclassification. In the case of non-differential misclassification (which is often the most pronounced since it is unrelated to the outcome of interest) this will in general result in attenuation of effects at least for binary exposures [28, 50]. In parallel, outcome misclassification will most often be non-differential since it is unrelated to the exposure of interest and probably closer related to the diagnostic process recorded in the register. The risk estimates in register-based studies are often small and non-differential misclassification could influence the final conclusion by removing significant effects due to this misclassification, e.g. for disease outcomes registered with low sensitivity an important association between exposure and outcome could be rejected because of non-differential misclassification.

In the Nordic countries, register-based epidemiological studies are based on exact linkage by the unique personal identification number of all residents in each country [4, 5]. The number is unique as it follows each person throughout her life and the same number is never given to a new person. In other countries without this unique personal identification number, methods have been developed to make probabilistic record linkage [51]. Studies have shown that these automatic methods give satisfactory results [52, 53], e.g. false-positive linkage rates of 2.2–4.7 % (over-inclusion) when linking emergency medical service data to hospital discharge data [52].

Finally, an investigator using registers for information on exposure or outcome should always make sure that all studied individuals are at risk at all times. Often it is necessary to choose persons who experience some event in the future (i.e. conditioning on the future), e.g. everyone included must survive for a specific period. Often this period of immortality comes from one of the entry criteria into the cohort. In this situation all analyses should exclude the period of risk time until entry criteria is met. If this immortal risk time is included the incidence rates will be underestimated because the denominator is inflated. Another example is the exclusion of persons unexposed before the occurrence of disease but exposed after disease occurrence. This may underestimate the risk of disease among unexposed persons thereby resulting in an upward biased risk estimate. The discussion of the paper by Dr. Kripke illustrates this bias [54].

Validity of registers

Validity of registers could be characterized as completeness (i.e. whether all individuals are included in the register), and validity of the variables included (i.e. whether all information on the persons are collected and whether the information registered is correct) [8].

The first dimension (completeness) refers to the proportion of individuals in the target population with the disease of interest, which is correctly included in the disease register. Completeness is closely related to sensitivity and positive predictive value (Table 1). Completeness could be evaluated by comparing the register of interest with another data source believed to be complete. This was done in a study of the Danish Registry on Regular Dialysis and Transplantation containing all Danish patients being actively treated for end-stage renal disease [55]. The register was linked with the Danish National Patient Register containing all admitted patients at Danish hospitals since 1977 [56] showing that the register had a completeness of 97.2 % [55].

Table 1 Relationship between terms in evaluation of validity disease outcome

	Disease status (gold standard)		
	Sick	Healthy	
Test outcome			
Test positive	True positive (A)	False positive (B)	Positive predictive value A/(A + B)
Test negative	False negative (C)	True negative (D)	Negative predictive value D/(C + D)
	Sensitivity A/(A + C)	Specificity D/(B + D)	Total

Another method is to compare the aggregated number of cases in a register with the total number in another data source, or, alternatively, to calculate the expected number of cases by applying rates from demographically similar populations. This was done in a study of toxic hepatitis where 512 cases were registered in the Danish National Patient Register from 1981 to 1985 [57]. During the decade 1978–1987 the Danish Board of Adverse Reactions to Drugs received 1,100 reports on hepatotoxicity. The authors concluded that these two figures were in close agreement thereby supporting an acceptable completeness of the Danish National Patient Register.

A third method is the capture–recapture method used to estimate the sensitivity of two case-finding methods [58]. The total number of cases can be calculated by assuming that the population of interest is closed (i.e. there is no change in the population during the investigation), that the presence of a case in the first sample is not influenced by the presence of the same case in the second sample, that cases sampled on both occasions can be identified and matched and each case has an equal chance of being included in each sample. This method has been used in several studies to estimate the prevalence of a disease and to evaluate the completeness of one data source [59, 60]. Alternative methods can be used to estimate the population size for open populations modeling capture heterogeneity [61].

A fourth method is to make a comprehensive patient chart review to evaluate whether all patients with a given disease is correctly classified into the register. This could be done in hospital discharge systems, where all cases should be registered, but some may be misclassified. This is a costly and time-consuming method, but is perhaps the most definite method when evaluating completeness of a particular register.

The second dimension is validity of the variables included and is the extent to which a variable measures what it is intended to measure. Important measures for validity are sensitivity, specificity and predictive value (both positive and negative) (Table 1).

Commonly, the validity of a register is performed by patient chart review of persons identified with a given disease in a register (case-to-case evaluation) and calculate the ratio between the number of correctly registered persons and all registered persons in the register [62–65]. This measures the positive predictive value of the registration (Table 1). An example is a study of persons with epilepsy diagnosis in the Danish National Patient Register including a random sample of 200 patients with an epilepsy diagnosis in the register and extracted information on several clinical characteristics from patient records [63]. One neurologist blinded towards the specific diagnosis of the patient classified the patients according to criteria. The authors found a positive predictive value of 81 % [63].

This method is valuable and sheds light on one dimension of registration validity, but does not contain information on true negatives. This means that when performing careful validity study as described above the incidence rate will be underestimated, because false negatives (C, Table 1) will be considered true negatives (D, Table 1).

The demand for high completeness and validity depends on the research question. In descriptive studies of disease frequency and in follow-up studies high completeness and validity are important, but in some analytical studies of association between exposure and outcome, high validity of outcome assessment is often more important than high completeness. In case–control studies it is more important that case ascertainment does not influence exposure classification than high completeness.

Conclusion

Despite intensive use of registers for research, a developed methodological literature is not available. We presented main strengths and limitations of registers for research. Furthermore, most important biases in register-based studies and methods for evaluating the validity of registers were discussed. We advance that register-based studies have important strengths compared to other data sources, but researchers should acknowledge the limitations and biases in epidemiological studies based on registers.

An important use of registers is linkage with research data, e.g. surveys. Hereby it is possible to add important background information and outcomes of interest from registers. Even if the missing data in a research study are not available there may be similar data or proxy measures in other registers. One important use of national register data is to study selection bias related to non-response.

Epidemiological studies with inclusion of all persons in a population followed for decades are important in modern epidemiology. The strengths of register-based studies are important to know and biases and limitations should also be highlighted for the scope of the results. In this paper we have presented the main strengths, limitations, biases and methods for evaluating validity of register-based studies.

References

1. Irgens LM, Bjerkeda T. Epidemiology of leprosy in Norway—history of National Leprosy Registry of Norway from 1856 until today. Int J Epidemiol. 1973;2(1):81–9.
2. Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. Epidemiol Rev. 1980;2:210–20.
3. St Sauver JL, Grossardt BR, Yawn BP, Melton LJ III, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. Am J Epidemiol. 2011;173:1059–68.
4. Olsen J, Bronnum-Hansen H, Gissler M, Hakama M, Hjern A, Kamper-Jorgensen F, et al. High-throughput epidemiology: combining existing data from the Nordic countries in health-related collaborative research. Scand J Public Health. 2010;38:777–9.
5. Thygesen LC, Daasnes C, Thaulow I, Bronnum-Hansen H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. Scand J Public Health. 2011;39:12–6.
6. Sorensen TI. Great scientific potential in Danish registries [in Danish]. Ugeskr Laeger. 1994;156:5812–3.
7. Frank L. Epidemiology—when an entire country is a cohort. Science. 2000;287:2398–9.
8. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol. 1996;25:435–42.
9. Sorensen H. Regional administrative health registries as a resource in clinical epidemiology. Aarhus: Aarhus University; 1996.
10. Sorensen H. Regional administrative health registries as a resource in clinical epidemiology. Int J Risk Saf Med. 1997;10:1–22.
11. Pike MC, Henderson BE, Casagrande JT, Rosario I, Gray GE. Oral-contraceptive use and early abortion as risk-factors for breast-cancer in young-women. Br J Cancer. 1981;43:72–6.
12. Brind J, Chinchilli VM, Severs WB, Summy-Long J. Induced abortion as an independent risk factor for breast cancer: a comprehensive review and meta-analysis. J Epidemiol Community Health. 1996;50:481–96.
13. Melbye M, Wohlfahrt J, Olsen JH, Frisch M, Westergaard T, Helweg-Larsen K, et al. Induced abortion and the risk of breast cancer. N Engl J Med. 1997;336:81–5.
14. Blenstrup LT, Knudsen LB. Danish registers on aspects of reproduction. Scand J Public Health. 2011;39(7 Suppl.):79–82.
15. Gjerstorff ML. The Danish cancer registry. Scand J Public Health. 2011;39(7 Suppl):42–5.
16. Norgaard M, Wogelius P, Pedersen L, Rothman KJ, Sorensen HT. Maternal use of oral contraceptives during early pregnancy and risk of hypospadias in male offspring. Urology. 2009;74:583–7.
17. Peltola M, Juntunen M, Hakkinen U, Rosenqvist G, Seppala TT, Sund R. A methodological approach for register-based evaluation of cost and outcomes in health care. Ann Med. 2011;43:S4–13.
18. Sund R, Nurmi-Luthje I, Luthje P, Tanninen S, Narinen A, Keskimaki I. Comparing properties of audit data and routinely collected register data in case of performance assessment of hip fracture in Finland. Methods Inf Med. 2007;46:558–66.
19. Dans PE. Looking for answers in all the wrong places. Ann Intern Med. 1993;119:855–7.
20. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med. 1988;318:352–5.
21. Irgens LM. Challenges to registry-based epidemiology in post-modernistic civilization. Nor Epidemiol. 2001;11:127–31.
22. United Nations Economic Commission of Europe. Register-based statistics in the Nordic countries. New York: United Nations; 2007.
23. Wallgren A, Wallgren B. Register-based statistics—administrative data for statistical purposes. Sussex: Wiley; 2007.

24. Hartley HO, Sielken RL Jr. A "super-population viewpoint" for finite population sampling. Biometrics. 1975;31:411–22.
25. Edington ES. Randomization tests. New York: Marcel Dekker; 1986.
26. Sorensen HT, Schulze S. Danish health registries. A valuable tool in medical research. Dan Med Bull. 1996;43:463.
27. Agerbo E. Epidemiological suicide research based on Danish routine registers. Aarhus: Aarhus University; 2009.
28. Olsen J. Register-based research: some methodological considerations. Scand J Public Health. 2011;39:225–9.
29. Jensen VM, Rasmussen AW. Danish education registers. Scand J Public Health. 2011;39(7 Suppl):91–4.
30. Olsen J. Using secondary data. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p. 481–91.
31. Thomsen CF, Skovdal J, Helkjaer PE. Intraobserver variation in the classification of diseases [in Danish]. Ugeskr Laeger. 1995;157:3746–9.
32. Green J, Wintfeld N. How accurate are hospital discharge data for evaluating effectiveness of care? Med Care. 1993;31:719–31.
33. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities. JAMA. 1988;260:2240–6.
34. Ray WA. Improving automated database studies. Epidemiology. 2011;22:302–4.
35. Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? Epidemiology. 2011;22:292–4.
36. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf. 2006;15:291–303.
37. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information. Epidemiology. 2005;16:17–24.
38. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol. 2000;29:722–9.
39. Hernan MA, Robins JM. Instruments for causal inference. An epidemiologists dream? Epidemiology. 2006;17:360–72.
40. Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. J Clin Oncol. 2001;19:1064–70.
41. Cavelaars AEJM, Kunst AE, Geurts JJM, Crialesi R, Grotvedt L, Helmert U, et al. Educational differences in smoking: international comparison. Br Med J. 2000;320:1102–7.
42. Groth MV, Fagt S, Brondsted L. Social determinants of dietary habits in Denmark. Eur J Clin Nutr. 2001;55:959–66.
43. Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. Int J Epidemiol. 2000;29:891–8.
44. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373–83.
45. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. 1992;45:613–9.
46. Ghali WA, Hall RE, Rosen AK, Ash AS, Moskowitz MA. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J Clin Epidemiol. 1996;49:273–8.
47. Clark DO, VonKorff M, Saunders K, Baluch WM, Simon GE. A chronic disease score with empirically derived weights. Med Care. 1995;33:783–95.
48. Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996;25:1107–16.
49. Groenwold RHH, Nelson DB, Nichol KL, Hoes AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. Int J Epidemiol. 2010;39:107–17.
50. Rothman KJ. Epidemiology—an introduction. Oxford: Oxford University Press; 2002.
51. Jaro MA. Probabilistic linkage of large public-health data files. Stat Med. 1995;14:491–8.
52. Dean JM, Vernon DD, Cook L, Nechodom P, Reading J, Suruda A. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. Ann Emerg Med. 2001;37:616–26.
53. Victor TW, Mera RM. Record linkage of health care insurance claims. J Am Med Inform Assoc. 2001;8:281–8.
54. Kripke DF, Langer RD, Kline LE. Hypnotics' association with mortality or cancer: a matched cohort study. BMJ Open. 2012;2:e000850.
55. Hommel K, Rasmussen S, Madsen M, Kamper AL. The Danish Registry on regular dialysis and transplantation: completeness and validity of incident patient registration. Nephrol Dial Transplant. 2010;25:947–51.
56. Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. Scand J Public Health. 2011;39(7 Suppl):30–3.
57. Almdal TP, Sorensen TI. Incidence of parenchymal liver diseases in Denmark, 1981 to 1985: analysis of hospitalization registry data. The Danish Association for the Study of the Liver. Hepatology. 1991;13:650–5.
58. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990–1993. The Clinical Epidemiology Group from Centres d'Information et de Soins de l'Immunodeficience Humaine. Int J Epidemiol. 2000;29:168–74.
59. Thomas AM, Thygerson SM, Merrill RM, Cook LJ. Identifying work-related motor vehicle crashes in multiple databases. Traffic Inj Prev. 2012;13:348–54.
60. Patterson CC, Gyurus E, Rosenbauer J, Cinek O, Neu A, Schober E, et al. Trends in childhood type 1 diabetes incidence in Europe during 1989–2008: evidence of non-uniformity over time in rates of increase. Diabetologia. 2012;55:2142–7.
61. McDonald TL, Amstrup SC. Estimation of population size using open capture-recapture models. J Agric Biol Environ Stat. 2001;6:206–20.
62. Devantier A, Kjer JJ. The national patient register—a research tool? Ugeskr Laeger. 1991;153:516–7.
63. Christensen J, Vestergaard M, Olsen J, Sidenius P. Validation of epilepsy diagnoses in the Danish National Hospital Register. Epilepsy Res. 2007;75:162–70.
64. Krarup LH, Boysen G, Janjua H, Prescott E, Truelsen T. Validity of stroke diagnoses in a National Register of Patients. Neuroepidemiology. 2007;28:150–4.
65. Djurhuus BD, Skytthe A, Faber CE. Validation of the cholesteatoma diagnosis in the Danish National Hospital Register. Dan Med Bull. 2010;57:A4159.