Methodological issues in using prescription and other databases in pharmacoepidemiology

Henrik Toft Sørensen[1], Søren Paaske Johnsen[2] and Bente Nørgård[1]

[1]*Department of Clinical Epidemiology, Aarhus University Hospital and Aalborg Hospital, Vennelyst Boulevard 6, DK-8000 Aarhus C, Denmark*
[2]*Department of Epidemiology and Social Medicine, University of Aarhus, DK-8000 Aarhus C, Denmark*

Correspondence: Henrik Toft Sørensen, Department of Clinical Epidemiology, Vennelyst Boulevard 6, Building 260, DK-8000 Aarhus C, Denmark
Telephone: +45 89 42 60 76 Telefax: +45 89 42 60 10 E-mail: hts@soci.au.dk

ABSTRACT

Development in technology has led to a considerable increase in the number of individual-based registers and databases, that may be of value in pharmacoepidemiological research, and the number of studies that are based on these secondary data may be expected to increase. The focus of this paper is to review methodological problems related to use of such databases in pharmacoepidemiological studies with respect to the four basic types of associations which can be observed in an observational study: 1) bias, 2) confounding, 3) chance or 4) causal. The following factors will affect the value and validity of registries and databases: 1) the completeness of registration of persons, 2) the validity and degree of completeness of the registered data, 3) the size of the data source, 4) the registration period, 5) data accessibility, availability, and costs, 6) data format, and 7) the possibilities of linkage with other data sources. The importance of these issues depends on the use of the data and on the problems they have to address. The Nordic countries have a unique possibility of record-linkage between registries because of the civil registry number assigned to every citizen at birth. In pharmacoepidemiological research this gives us the opportunity to study different outcome events in relation to drug use, and this has been extensively used in the Danish pharmacoepidemiological approach. The Nordic countries could play a leading role in future pharmacoepidemiological research. This, however, requires considerably more efficient and comprehensive use of the collected data on which the society has spent many resources for other purposes.

INTRODUCTION

The randomised controlled trial (RCT) is the best way of studying drug efficacy with its potential for controlling the influence of confounding variables. With respect to effectiveness and side effects, this design has many limitations because 1) the outcomes are most often too rare, 2) ethical barriers, 3) standardised interventions may be different from common practice, 4) RCTs tend to restrict the scope and narrow the study question, 5) RCTs are costly in time and money, and 6) RCTs often have too short follow-up time to detect side effects. Therefore, observational studies, as cohort and case-control designs, have been introduced in the clinical pharmacology to study the use of and effects of drugs in a large number of people (1). This discipline, bridging clinical pharmacology and epidemiology, is called pharmacoepidemiology. Since pharmacoepidemiological studies need to be large to study effectiveness and rare adverse effects, they are often conducted on existing databases covering many exposed individuals. The focus of this paper is to review methodological problems related to use of such databases with respect to the four basic types of associations which can be observed in an observational study: 1) bias, 2) confounding, 3) chance or 4) causal.

Among the examples of adverse and possible beneficial effects of drugs which have been revealed by systematic collection of data are the positive association between thalidomide and phocomelia (2), oral contraceptives and venous thrombotic episodes (3), the negative association between statins and dementia (4), and antibiotics and acute myocardial infarction (5). Furthermore, pharmacoepidemiological studies may also provide useful information on patterns of drug utilization and pharmacoeconomic issues.

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

Limited methodological knowledge

Despite the increasing use of administrative databases in research, little has been published on critical issues involved in this type of research (7-11). Epidemiological research issues in particular raise questions that have not received adequate attention in the literature. The present review is based on our research experience within pharmacoepidemiology with administrative registries in Denmark and the international literature.

Historical perspective

The Nordic countries have for many years established a number of disease and administrative registries. In Denmark all births and deaths have been registered in church files since 1645, and in 1769 the first census was taken (6). The first disease registry was started as early as 1856 with the Leprosy Registry in Norway. In the present century registries on causes of death, tuberculosis, and cancer (in Denmark in 1943) were added. An important milestone in Danish pharmacoepidemiology was the establishment of the two Prescription Registries in 1989/90 in Funen and North Jutland Counties. The establishment of the National Population Registry in 1924, and the civil registry number (the CPR-number) in 1968, allowed for person-identification of a remarkable quality, and for the possibility of collecting information about the same person in several independent registries, which has been used in pharmacoepidemiology in Denmark where outcomes have been collected in other registries (6). This is a rather unique situation for the Nordic countries compared with that of other countries.

The computerization of several administrative registries has, however, also increased the use of administrative data sources in pharmacoepidemiology outside the Nordic countries. This has occurred particularly in Canada and the USA, where data from MEDICARE, MEDICAID, Harvard Community Health Plan, Kaiser Permanent Medical Care Program, and health insurances in Manitoba and Saskatchewan (9-12) have been used. In Europe, the General Practice Research Database in the UK (13) together with databases in Scotland (14), the Netherlands, and Italy have become valuable tools within pharmacoepidemiological research.

Use of existing registries

The concepts of originality and credibility are essential in all medical research, whether registry-based or not. Any study based on secondary data should be designed with the same critical approach as with studies based on primary data, i.e. specifying hypotheses, estimating sample size to obtain valid answers, and aiming at reducing both systematic (bias) and random errors with the aim of distinguishing between the four types of associations (6,15). In this way, registry-based research is not different from other types of research.

Using existing registries in pharmacoepidemiological research, the following factors (Table 1) will affect the value and validity of registries and databases (15): 1) the completeness of registration of persons, 2) the validity and degree of completeness of the registered data, 3) the size of the data source, 4) the registration period, 5) data accessibility, availability, and costs, 6) data format, and 7) the possibilities of linkage with other data sources. The importance of these issues depends on the use of the data and on the problems they have to address.

Table 1. Factors affecting the value of registry data in pharmacoepidemiological research.

1. Completeness of registration of persons
   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

The completeness of registration of persons

The use of existing registries in medical research is due to many advantages (6,15). One of these is the frequent completeness of the registries with respect to the people in the target population (6,15), which ensures representativeness. However, 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 (6,15). 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 (15). 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.

The validity of registered data will often have to be evaluated by comparison with independent external criteria (15), and with respect to completeness and validity, the prescriptions and the diagnoses will often have to be compared with operational criteria by going through records (12).

The accuracy and degree of completeness of variables

When registries are used for pharmacoepidemiological research, information on the drug exposure should of course, if possible, be well-defined in terms of timing, dose, and duration of use. Furthermore, the available information about exposures to the drug must be accurate and complete, and the exposure definition must be relevant for the question under study. For instance, for many outcomes with long latency period as cancer, data about dose and duration of use are essential.

One of the limitations in using administrative registries for research is related to data selection and quality, since the methods of data collection are predetermined, not controlled by the researcher, and sometimes impossible to validate (6,15). A number of studies have reported high validity estimates of diagnoses frequently used as outcome within pharmacoepidemiology, e.g. venous thromboembolism (16) and gastrointestinal bleeding (17). The validity may, however, vary between different settings, and validation studies are therefore important when using registries as a datasource. It is well known that use of discharge diagnoses for the identification of cases can cause problems, since at least five sources of error have been described: a) variation in coding procedures, b) coding errors, c) incomplete coding, d) lack of specificity in available codes, and e) error in the clinical diagnosis (18). Misclassification of outcomes in the range of 10 to 30 percent is often seen, and will bias the risk estimates towards the null if the misclassification is random (15).

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. When evaluating variables one also has to consider the extent of missing data, since a significant degree of missing and incomplete data negates the value of the source (7,19). For each single variable it should be considered whether missing information means that exposure or outcome have 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 (20).

The size of the data source

One of the advantages of registries is their large size, which allows for great statistical precision of estimates and makes it possible to study rare exposures, diseases, and outcomes (6,15).

When using registry data in research, it is of course essential to know how many persons 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 dimensions (15). At the same time one must remember that using restriction or 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.

Registration period

Often data sources only contain cross-sectional registrations, 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. For instance, data sources with observation periods of a few years will seldom be suitable for drug induced cancer research. Another aspect of registration period is the considerations on the timing of events in relation to start of drug exposure. One should remember that in studies of adverse effects of a drug, it may be useful to be able to distinguish between new users and past users, because past use of a drug, and chronic prescribing will tend to be associated with nonsusceptibility to adverse effect of the drug (1). Therefore, it may not be desirable to mix together subjects with different patterns of drug usage.

Data accessibility and availability

It is often not clear who owns the data and who has the right to use them (accessibility) (21). It is important to clarify these points and to find out which authorities should approve the use of the data for research purposes. 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 registered persons cannot reach unauthorized third parties.

Data format

During a certain registration period, codes, and even the layout of records, are often changed periodically. These changes in codes, diagnostic criteria and classifications (e.g. the recent change to ICD-10 disease classification system and changes in ATC-codes) frequently cause problems when comparing data over longer periods (15).

Record-linkage data sources

In Denmark (and the other Nordic countries) we have a unique possibility of record-linkage between registries because of the CPR-number assigned to every citizen at birth. In pharmacoepidemiological research it gives us the opportunity to study different outcome events, e.g. cancer, other diseases, and birth outcome in relation to drug use (6,15). This has been extensively used in the Danish pharmacoepidemiological approach.

THE DANISH PHARMACO-EPIDEMIOLOGICAL APPROACH FOR ANALYTICAL STUDIES

A pharmacoepidemiological model has been developed based on existing data sources in Denmark (Figure 1). All pharmacies in Denmark use databases in connection with the accounting of the prescriptions in the National Health Service (22). The accounting system provides no information about measurement of drug effects other than prescriptions of other drugs, but this can be obtained by record linkage between different outcome registries, e.g. the Hospital Discharge Registry, the Birth Registry, the Death files, and the Cancer Registry. Regarding the quality of information in these registries the validity of the diagnoses in the Hospital Discharge Registry is varying between 14% and 100% depending on the diagnoses (23), the quality of the data in the Birth Registry is reportedly good for birth weight, birth complications, and parity (24,25), and the completeness and validity of information in the Cancer Registry is 95-98% (26). Within the pharmacoepidemiological model, it is possible to combine information from different data sources in a number ways (27-30) as described in the following examples:

1) In a population-based cohort study the incidence of cancer after antidepressant treatments was examined (27). The Pharmaco-Epidemiologic Prescription Database of the County of North Jutland, Denmark, was used to identify 39,807 adult users of antidepressants, and information on cancer occurrence was obtained by linkage to the Danish Cancer Registry. The number of cancers among users of antidepressants was compared with the number that would be expected, on the basis of age-, sex-, and calendar year-specific incidence rates of first primary cancer in the population of North Jutland. In the follow-up period beginning one year after first known prescription, there were 766 cancers among users of antidepressants compared with 746 expected, for a standardised incidence rate ratio (SIR) of 1.0 (95% CI = 1.0–1.1). Thus, there was no overall increase in cancer risk among individuals taking antidepressant medication, but among regular users of tricyclic antidepressants an increased risk of non-Hodgkin’s lymphoma was observed (SIR = 2.5; 95% CI = 1.4–4.2).

2) In a record linkage study between the same Pharmaco-Epidemiologic Prescription Database and the Hospital Discharge Registry, we examined the risk of hospitalisation for upper gastrointestinal (GI) bleeding with use of low-dose aspirin (29). Incidence rates of upper GI bleeding were compared with the incidence rates in the general population. A total of 207 exclusive users of low-dose aspirin experienced a

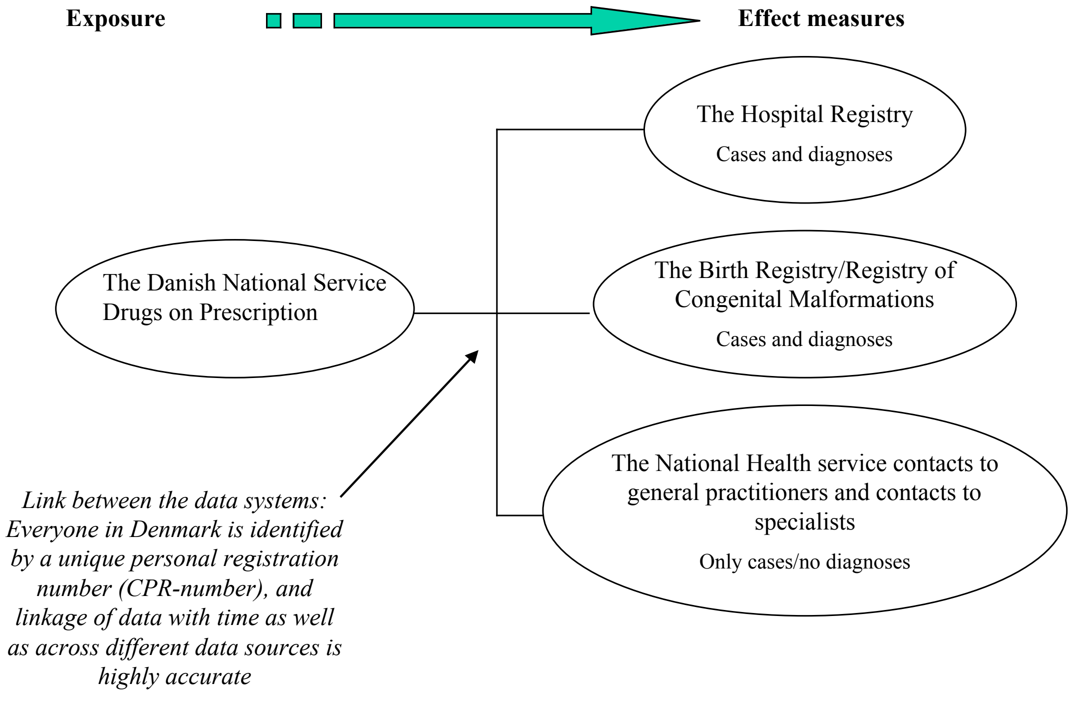
Exposure
Effect measures
The Hospital Registry
Cases and diagnoses
The Danish National Service Drugs on Prescription
The Birth Registry/Registry of Congenital Malformations
Cases and diagnoses
The National Health service contacts to general practitioners and contacts to specialists
Only cases/no diagnoses
Link between the data systems: Everyone in Denmark is identified by a unique personal registration number (CPR-number), and linkage of data with time as well as across different data sources is highly accurate

Figure 1. A pharmacoepidemiological analytical model based on existing data sources in Denmark.

first episode of upper gastrointestinal bleeding during the study period. The SIR was 2.6 (95% CI = 2.2–2.9). The risk was similar among users of noncoated and coated low-dose aspirin. These findings raise the possibility that prophylactic use of low-dose aspirin may convey an increased risk of gastrointestinal bleeding, which may offset some of its benefits.

3) Based on record linkage between the Pharmaco-Epidemiologic Prescription Database, the Danish Cancer Registry, and the Mortality files, the cancer risk and mortality among 23,167 users of calcium channel blockers in North Jutland County, Denmark, was examined (30). Overall, 967 incident cancers occurred, resulting in a SIR of 1.04 (95% CI = 0.98-1.11). There was a slightly elevated non-significant risk of tobacco-related cancer. No increased risk of breast or colon carcinoma was observed. The cancer mortality was close to that expected in the background population (standardised mortality ratio of 0.97; 95% CI = 0.89–1.04). This large-scale, population-based study adds to the increasing evidence indicating no substantial association between the use of calcium channel blockers and the incidence rate of cancer or cancer mortality.

4) In another study (28) the risk of adverse pregnancy outcome in women exposed to nonsteroid anti-inflammatory drugs was examined in population-based follow-up and case-control studies based on drug exposure data from the Pharmaco-Epidemiologic Prescription Database and outcome data from the Danish Birth Registry and the Hospital Discharge Registry. In the cohort analysis a total of 1462 women with a live birth or stillbirth after 28 gestational week had redeemed 1,742 prescriptions for nonsteroid anti-inflammatory drugs. As reference cohort was selected all pregnancies in the same period in which the mothers had no prescription of any kind (17,269 women). The odds ratios of congenital abnormalities, low birth weight, and preterm birth were 1.27 (95% CI = 0.93–1.75), 0.79 (95% CI = 0.45–1.38), and 1.05 (95% CI = 0.8–1.39), respectively. In the case-control analysis nonsteroid anti-inflammatory drug exposure in women having a spontaneous abortion was compared with pregnancies ending with a birth. Odds ratio of spontaneous abortion was 6.99 (95% CI = 2.75-17.74) when nonsteroid anti-inflammatory drugs were used in the last week before abortion. Thus, the use of nonsteroid anti-inflammatory drugs during pregnancy does not seem to substantially increase the risk of adverse birth outcome, but is associated with risk of spontaneous abortion.

CONCLUSIONS

The health registries contain several data that provide relevant information for clinical and epidemiological research. The registries cover essential aspects of disease, utilization of care, and its determinants (6). Danish prescription registries and other registries clearly offer important advantages and have a high research value at an international level because of the free access to health care and the regional and nationwide primary and secondary health care data files. Compared with international data sources, they cover populations of persons whose numbers are known, and, in longitudinal studies, they give a low rate of loss.

However, several problems remain. As with all available registries, there are limitations with the available information, and it is important that they are understood (6,15). Secondary data do not cover all aspects of interests, and the presently available data seldom include information about confounding variables, such as indications, lifestyle factors, and other risk factors. Therefore, in some cases information must be supplemented by questionnaires and use of medical records. The use of registries in research depends on their accessibility and quality, and quality is linked with their use. Therefore, accessibility and quality should be improved, and furthermore, the registries should be improved basically to benefit research needs. The main problem connected with the use of hospital registries is the great variation in coding practices and quality, as well as the applicability of the disease classification. Finally, as unexpected findings appear especially often in pharmacoepidemiological studies, where multiple outcomes are commonly sought in a population taking many drugs, the risk of finding spurious associations should not be overlooked. Thus, unsual results should always be interpreted with caution since they may be a play of chance, but may also represent the first indication of a beneficial or adverse effect of the studied drug (31).

The Nordic countries could play a leading role in future pharmacoepidemiological research. This, however, requires considerably more efficient and comprehensive use of the collected data on which the society has spent many resources for other purposes. The challenges presented are thus 1) to include a number of ordinary diseases in registry-based research, 2) to improve data quality, 3) to spread knowledge about epidemiological methods, and 4) to make more frequent use of the registries in clinical and public health research.

REFERENCES

1. Strom BL (ed). *Pharmacoepidemiology*. New York: Churchill Livingstone, 1989.
2. McBride WG. Thalidomide and congenital abnormalities. *Lancet* 1961; **ii**: 1358.

3. Andersen BS, Olsen J, Nielsen GL, Steffensen FH, Sørensen HT, Baech J, Gregersen H. Third generation oral contraceptives and heritable thrombophilia as risk factors of non-fatal venous thromboembolism. *Thromb Haemost* 1998; **79**: 28-31.
4. Jick H, Zornberg GL, Jick SS, Seshadri S, Drachman DA. Statins and the risk of dementia. *Lancet* 2000; **356**: 1627-31.
5. Meier CR, Derby SS, Jick SS, Vasilakis C, Jick H. Antibiotics and risk of subsequent first-time acute myocardial infarction. *JAMA* 1999; **281**: 427-31.
6. Sørensen HT. Regional administrative health registries as a resource in clinical epidemiology. *Int J Risk Safety Med* 1997; **10**: 1-22.
7. Connel FA, Diehr P, Hart LG. The use of large data bases in health care studies. *Annu Rev Public Health* 1987; **8**: 51-74.
8. Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. *Epidemiol Rev* 1980; **2**: 210-20.
9. Lauderdale DS, Furner SE, Miles TP, Goldberg J. Epidemiologic uses of Medicare data. *Epidemiol Rev* 1993; **15**: 319-27.
10. Bright RA, Avorn J, Everitt DE. MEDICAID data as a resource for epidemiologic studies: strengths and limitations. *J Clin Epidemiol* 1989; **42**: 937-45.
11. Ray WA, Griffin MR. Use of MEDICAID data for pharmacoepidemiology. *Am J Epidemiol* 1989; **129**: 837-49.
12. 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.
13. Garcia Rodríguez LA, Gutthann SP. Use of the General Practice Research Database for Pharmacoepidemiology. *Br J Clin Pharmacol* 1998; **45**: 419-25.
14. Evans JMM, MacDonald TM. Record-linkage for pharmacovigilance in Scotland. *Br J Clin Pharmacol* 1999; **47**; 105-110.
15. Sørensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol* 1996; **25**: 435-442.
16. Lawrenson R, Todd J-C, Leydon GM, Williams TJ, Farmer RDT. Validation of the diagnosis of venous thromboembolism in general practice database studies. *Br J Clin Pharmacol* 2000; **49**: 591-6.
17. Cattaruzzi C, Troncon MG, Agostinis L, Garcia Rodríguez LA. Positive predictive value of ICD-9th codes for upper gastrointestinal bleeding and perforation in the Sistema Informativo Sanitario Regionale Database. *J Clin Epidemiol* 1999; **52**: 499-502.
18. Steinberg EP, Whittle J, Anderson GF. Impact of claims data research on clinical practice. *Int J Tech Assess Health Care* 1990; **6**: 282-7.
19. Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. *JAMA* 1992; **267**: 2197-203.
20. Rothman KJ. *Modern epidemiology*. Boston: Little Brown and Company, 1986.
21. Smith GD. Increasing the accessibility of data. *BMJ* 1994; **308**: 1519-20.
22. Sørensen HT, Larsen BO. A population-based Danish data resource with possible high validity in pharmacoepidemiological research. *J Med System* 1994; **18**: 33-8.
23. Devantier A, Kjer JJ. The national patient register – a research tool. *Ugeskr Laeger* 1991; **153**: 516-7.
24. Knudsen LB. Information on parity in the medical registry of births of the National Board of Health. *Ugeskr Laeger* 1993; **155**: 2525-9.
25. Kristensen J, Langhoff Roos J, Skovgaard LT, Kristensen FB. Validation of the Danish Birth Registration. *J Clin Epidemiol* 1996; **49**: 893-7.

26 Storm HH, Michelsen EV, Clemmensen IH, Pihl J. The Danish Cancer Registry – history, content, quality and use. *Dan Med Bull* 1997; **44**: 549-53.

27. Dalton SO, Johanssen C, Mellemkjær L, Sorensen HT, McLaughlin JK, Olsen J, Olsen JH. Antidepressant medications and risk for cancer. *Epidemiology* 2000; **11**: 171-6.
28. Nielsen GL, Sørensen HT, Larsen H, Pedersen L. Risk of adverse birth outcome and miscarriage in pregnant users of non-steroidal anti-inflammatory drugs: population based observational study and case-control study. *BMJ* 2001; **322**: 266-70.
29. Sørensen HT, Mellemkjær L, Blot WJ, Nielsen GL, Steffensen FH, McLaughlin JK, Olsen JH. Risk of upper gastrointestinal bleeding associated with use of low-dose aspirin. *Am J Gastroenterol* 2000; **95**: 2218-24.
30. Sørensen HT, Olsen JH, Mellemkjær L, Thulstrup AM, Steffensen FH, McLaughlin JK, Baron JA. Cancer risk and mortality in users of calcium channel blockers. *Cancer* 2000; **89**: 165-70.
31. Olsen JH. Interpretation in drug epidemiology. *Lancet* 1998; **352**: 162-3.