

The use of a national hospital register for hypothesis generation

Finn Tüchsen and Elsa Bach ^a
in collaboration with Otto Andersen ^b and Jørgen Jørgensen ^c

^a *Danish National Institute of Occupational Health, 105 Lersø Park Allé, 2100 Copenhagen Ø, Denmark*

^b *Danmarks Statistik*

^c *The National Board of Health*

Administrative data collected in large registers such as the Danish National Inpatient Register can be used to generate and even test hypotheses on associations between environmental health hazards and hospitalization. This paper describes the necessary record linkages and transformations before an outcome file can be made available in a database. Special attention is drawn to the question of confidentiality.

Introduction

In modern societies we have a considerable number of large administrative files. These files can, to some extent, be used for research but usually not without costly transformations and linkage with other files. Despite the fact that cause of death data has been filed for centuries, England and Wales are the only countries in which regular occupational mortality statistics have been produced until a couple of decades ago. Several countries have, however, recently constructed very informative surveillance systems on variations in occupational mortality. Some valuable local systems on hospitalization have been made [1,2,3] but no nationwide system has yet been set up.

Denmark has a unique nation-wide registration of all hospitalizations, but this National Inpatient Register can not, as such, be used for generating of hypotheses on environmental health risks and hospitalization.

In order to be able to analyze associations between occupation and hospitalization, a new research register had to be established through the linkage of several administrative registers. This paper deals with the linking procedures and the first results of analysis of “The Occupational Hospitalization Register”. Further, how such a research register can be used for generation and testing of hypothesis is discussed.

An outcome register containing all combinations of diagnosis, industry and occupation could be a valuable part of an Environmental Health Information System (EHIS).

The Occupational Hospitalization Register has been set up with several purposes. The benefit is expected to be (1) knowledge of occupational differences in hospitalization by selected diagnosis, (2) development of methods and programs for a permanent information system, (3) validation of the basic information, and (4) formation of a cohort for in-depth analysis.

Confidentiality

As information in this study includes diagnosis on people still alive, confidentiality has to be absolutely secured. Criminal misuse must, as far as possible, be hindered.

The Public Registers Act in Denmark includes provisions on several aspects of confidentiality. This study has been approved by The Register Inspection. The Register Inspection must secure that registers are deleted after a few years, unless permissions are prolonged. Public registers must not include unnecessary information. The law of administration from 1987 incriminates misuse of information held by public servants, including the staff of this study.

This study has been approved by The Register Inspection, the Minister of the Interior, the Minister of Economic Affairs, the Minister of Labour, and a council of independent physicians associated with the National Inpatient Register. The most important precaution is, however, that the information included has been restricted to what is absolutely essential in regard to the purpose of the study.

There are technical security issues in handling and computing of data. Information about these systems is not published.

To prevent that any individual person can be traced by combining information on gender, age, occupation, industry etc., only results including at least 20 persons in an occupational group, and at least 5 observed hospitalizations per age group, will be published. If it is decided to publish the results in a database, groups will be aggregated until these two conditions are fulfilled.

Hospitalization versus mortality

The occupational mortality surveillance systems provide valuable information about mortal diseases, but information about non-mortal diseases are not available on a regular basis. As the progress of different mortal diseases are very unlike, incidence and prevalence of the disease can not be estimated from mortality registers. Furthermore, many diseases treated in hospitals are usually not mortal. Part of the purpose of this study is to analyze the possibilities of establishing reliable incidence measures from hospital statistics.

Definitions

An age-specific rate usually refers to the number of deaths or diseases in a defined age group, where the denominator is the average or midyear population. In the many Nordic cross-linked mortality studies, the population at the begin-

ning of the period is used as denominator. We have chosen to follow the Nordic standard. A consequence of this is, that the incidences, we report, are usually relatively higher than the incidences in international statistics.

In the National Inpatient Register, the National Board of Health use the words “admittance” and “discharge” regarding a single department and a single episode. In accordance with the recommendations of Christau et al. [4] we use the term “administrative pass” to describe the period a patient is “admitted” to or transferred to a specific department, until he is “discharged” to home or another department.

Material

The sources of information in this study are the Central Population Register, the Employment Classification Pool, and the National Inpatient Register, the Migration Files, the Death Certificate Register and the Population and Dwelling Register.

The Central Population Register

To be able to calculate the period of risk we have entered some additional information. From The Central Populations Register we have included information on The Personal Identification Number, from which we extract the information about gender, date and year of birth. The Central Population Register enables us to identify the 3,327,758 persons actually living in Denmark on January 1st 1981 aged 15 to 64 years (at present only people aged 20 to 59 years are represented in the actual linked file and in the outcome file).

The Employment Classification Module

The census in November 1970 was the last traditional census in Denmark. Since then censuses and employment-classifications have been made every year, based on information collected from different administrative registers. The classifications are organized in the Employment Classification Pool (ECP). The ECP contains information on economic activity, occupational classification and industry for each person in Denmark aged 16+ years.

- (1) Economic activity is a variable dividing the population in eight groups: (1) self-employed, (2) assisting wife, (3) employed with identification of self-employment, (4) employed, and (5) employed with limited income, (6) retired with identification of self-employment, (7) retired, and (8) others.
- (2) Occupational classification. The salary earners are coded as such, while skilled wage earners, unskilled wage earners and employees not further specified are divided in occupational groups according to a special “Danish Occupational Code”, which is very close to the 1968 version of “International Standard Classification of Occupations” (ISCO). The original ISCO holds 55

categories, but it has been extended and holds 295 categories in the version used by Danmarks Statistik.

- (3) Industry is coded in accordance with the 1968 version of “International Standard Industrial Classification of All Economic Activities” (ISIC). The first five digits are equal to ISIC, while the last digit represents a national extension. Industry is divided into almost 600 categories. The categorization is applied to local production plants, rather than to head offices of the companies.

The National Inpatient Register

The National Inpatient Register holds information on admission to Danish hospitals. The register collects information from on-going register systems in each of the 16 county registers in Denmark. The National Inpatient Register is updated and stored in a separate volume each year. From 1980 all admissions to Danish Hospitals are covered. The 1981 to 1984 volumes are included in this study. The information we pick out from the National Inpatient Register is the Personal Identification Number, which is used in the cross-linking procedure, but excluded in the final file. Further, we use the date of “admission”, the date of “discharge”, all the discharge diagnosis and modifications which count up to a number of twenty. For the age groups 15–64 years, The National Inpatient Register held 2,140,315 administrative passes for the four included years, after 9,443 have been excluded due to irrelevance. The excluded numbers are constructed numbers for foreigners treated at Danish hospitals.

Supplementary information

Death and emigration in the follow-up period, January 1st 1981 to December 31st 1984, were identified by linkage with the Death Certificate Files and the Migration files from 1981 to 1984.

People are no longer at risk to be admitted to a hospital in Denmark, from date of first emigration (59,973 persons) or from date of death (59,566 persons). From the date of the event these persons are excluded from the denominator in the risk calculations. This information is collected from four years of the migration files and the death certificate files.

Validity of basic information

The validity of administrative registers are mainly of two kinds:

- (1) information which is checked by the authorities, because it is important for administrative purposes, and
- (2) supplementary information, which is not checked because it is not used in the administrative processes.

The basic register holds both kinds of information. If information is available from different sources, the most reliable have been chosen. Knowledge of the

origin of data should be included in the guiding expert system if the table files are made available for public use.

The Employment Classification Module has not been validated so far and the National Inpatient Register has not been validated on a regular or nation-wide base, but a pilot study including a single hospital has been done. This hospital is not representative for all Danish hospitals [5]. An up-to-date validation would be very desirable.

Methods

The Occupational Hospitalization Register has been set up by linking the Central Population Register, the Employment Classification Module and the National Inpatient Register. The Employment Classification Module has been set up by a linking process including a row of administrative registers.

Linking procedures

The linking procedures include 10 single procedures.

- (1) From the Central Population File all persons aged 15 to 64 years are drawn per January 1st 1981. Persons born between January 2nd 1926 and January 1st 1966, both dates included, are in the cohort.
- (2) From the Dwelling and Housing Register information on urbanization is included. Two kinds of urbanization codes are included.
- (3) The population as of January 1st 1981 is sorted by the Personal Identification Number.
- (4) The study population is defined. Persons found in the Central Personal Identification Register, but not in the Employment Classification Module, are classified with unknown occupation.
- (5) The final study population is defined as persons with employment status 1 to 5 and persons undefined in the former procedure.
- (6) The relevant age group is picked out from the death certificate files and sorted before
- (7) the dates of death are included in the relevant records.
- (8) Date of (first) emigration or disappearance are included from the Migration Files.
- (9) The National Inpatient Register is cleaned for incorrect Civil Identification Numbers (these are constructed numbers for foreigners admitted to Danish hospitals). 9,443 records (administrative pass) were deleted. From the migrations files we include date of first emigration and from the death certificate files we get the date of death.
- (10) Finally we add from the National Inpatient Register, all discharge diagnoses and modifications, date of admittance and discharge (for each administrative pass).

We have hereby formed the Occupational Hospitalization Register in its first rough version. What we have now is a cohort including all persons in Denmark aged 15 to 64 years on January 1st 1981.

The next aim is to establish a register for analysis. As less than half of the persons in the age groups 15–19 and 60–64 years (as of January 1st 1981) can be related to an occupation potentially associated to their hospitalization, the final linked register is restricted to the age groups 20–59 years. At this step 1,293,888 men and 1,285,508 women are enrolled in the study.

The National Inpatient Register holds information on persons admitted to hospital for observation or with a suspected diagnosis which was not confirmed. Administrative passes with modification 1 (63,454) or 2 (11,002) on the principal diagnosis are not included in the final register. We have also chosen to exclude admissions due to normal births and diseases of pregnancy (150,178). Finally admissions for examination, “Y-diagnosis” as the principal diagnosis (94,725 administrative passes, mainly related to pregnancy), are excluded.

Establishing and algorithm of admissions

The National Inpatient register is primarily established for administrative purposes, so an “admission” is defined as an admission to a department (which in our terminology is an administrative pass). When a person is transferred to another department, it is filled as two independent admissions. Calculation of true incidence rates is therefore not possible unless the real duration of the disease can be reconstructed.

We have three possibilities and a combination of these to do so. (1) The code “Admitted from” and the code “Discharged to” gives a hint of whether the patient has finished his hospitalization, or if he is transferred to another department, (2) the number of nights from the discharge to next admission, and (3) the diagnosis. The pilot study has demonstrated that none of these possibilities are satisfactory.

In consequence we have chosen to measure the relative risks in two age standardized measures: (1) observed to expected percentage admitted persons with the relevant diagnosis as the principal diagnosis, and (2) observed to expected percentage of persons with the relevant diagnosis mentioned as one of all diagnoses.

To make the measures sensitive it is essential that period of risk is calculated exactly and for each person and each diagnosis separately. This step is costly in both manpower and CPU time, but absolutely necessary when we want the register to uncover more than it hides.

Finally we are able to produce the tables which could be included in a database, accessible on a host in an Environmental Health Information System or distributed as a separate database, on a compact disk.

Examples

Hypotheses could have the form of a single table like Table 1. This tabulation raises the hypothesis that females working in the nutrition and stimulant in-

Table 1

4,677 unskilled female workers in nutrition and stimulant industries. Principal diagnoses ICD-8: 412-4, 440. Chronic ischaemia and atherosclerosis

Age 1981	No. of persons	Years at risk	Obs. No.	Exp. No.	O/E × 100
20-44	8,039	31,917	13	5.25	248
45-49	1,638	6,445	13	6.15	211
50-54	1,422	5,570	18	8.05	224
55-59	1,282	5,016	14	9.39	149
20-59	14008	55413	59	29.12	203

Table 2

4,690 unskilled male workers in rubber and plastic industry. Principal diagnoses ICD-8: 412-4, 440. Chronic ischaemia and atherosclerosis

Age 1981	No. of persons	Years at risk	Obs. No.	Exp. No.	O/E × 100
20-59	1,628	6,274	11	7.78	141

dustries have an excess risk of being admitted to hospital due to chronic ischemia and atherosclerosis.

A well-known cardiotoxic pollutant in the rubber industry and in viscose rayon plants is carbon disulfide. In a recent Danish study [6] 0.17-0.7 mg/m³ air was measured in the breathing zone of vulcanization workers. The excess hospitalization for unskilled male workers in rubber and plastic plants due to chronic ischemia and atherosclerosis was therefore to be expected. (See Table 2.)

A surprising but consistent result of the occupational mortality statistics has been a high mortality for all persons occupied in hotels and restaurants (SMR = 155 for males in all Nordic Countries together [7]). In Denmark male servants have an excess mortality from heart diseases (SMR = 150) [8]. We can therefore use The Occupational Hospitalization Register to test the hypothesis that working in hotels and restaurants is associated with an excess risk of getting hospitalized due to ischemic or atherosclerotic diseases. (See Table 3.)

Table 3

Hospitalized men working in hotels and restaurants 1981. Main diagnosis ICD-8: 412-14 and 440. Chronic ischaemia and atherosclerosis

Occupational group	O	E	O/E
1,320 + 1,322 + 2,320 self-employed and assisting spouses	64	35.13	182
4,351 managers	25	10.91	229
4,451 other salary earners	17	7.67	222
4,551 skilled waiters and cooks	36	21.32	169
4,651 unskilled workers	18	10.15	177
All men in hotels and restaurants	160	85.18	188

Table 4

Hospitalized women working in hotels and restaurants in 1981. Main diagnosis ICD-8: 412-14 and 440. Chronic ischaemia and atherosclerosis

Occupational group	O	E	O/E
1,320 + 1,322 self-employed	21	8.76	412
2,320 assisting wives	9	4.69	192
4,351 managers	6	1.95	308
4,451 other salary earners	10	7.31	137
4,551 skilled waiters and cooks	5	1.98	253
4,651 unskilled workers	278	206.35	135
All women in hotels and restaurants	329	231.04	142

In the Occupational Mortality Statistics waitresses and sandwich makers also had an excess heart mortality (SMR = 136). The results in Table 4 support these findings.

Discussion

Research registers like the “Occupational Hospitalization Register” are often used as a monitoring system or for hypothesis generation. Up to now mortality registers have often been used as reference material for testing of hypotheses.

In our opinion epidemiological research is a never ending attempt to improve data, to analyze it in better ways, to describe, measure and avoid bias, to compare results obtained by other methods with the epidemiological results. So, we hold the view that this register is one step forward. The main advantage is that a very large part of the population is analyzed in a comparable way. The large number of persons and episodes included has made it possible to use more homogeneous groups than usual. This fact has improved the power of the study. Therefore the results are more similar to ad hoc studies than former register-based studies.

The next step could be that additional information on occupation is included, so that primary and secondary “healthy worker” selection can be controlled. More precise measures of, or at least estimates of, exposure should be included from other sources such as our measuring databases and the databases of products and chemicals. The register should be regularly up-dated with new information on hospitalization and new information on occupation. The register should also be up-dated with new persons. This will especially facilitate analysis of diseases with long latency periods, as it will facilitate both cohort and dynamic population studies.

Analyses on selected topics should be carried out selecting and comparing more homogeneous groups than have been possible in the first overall description. Knowledge of exposure and absence of exposure is needed in order to select reasonable study and reference groups. Easy access to toxicological databases in an Environmental Health Information System could be helpful in this respect.

Conclusion

If all occupational data in the Danish Occupational Hospitalization Register are tabulated and entered into a database, the total size is at least 1 Gbyte or 0.5 mio., A4 pages. If information on demographic, housing and unemployment variables are included the total size can be multiplied by a factor of 10 or more. The only reasonable distribution form is a database. It could be distributed as a CD-ROM or included in an Environmental Health Information System. In both cases researchers in other countries will have access to a very cheap and fast possibility of launching and even testing hypothesis concerning relations between occupations and industries on one side and hospitalizations due to the diagnosis in the WHO classification on the other. All data have been collected, linking procedures have been established, calculations of individual risk time and standardization have already been done for some diagnosis, but much work still has to be done in order to make the database accessible from various countries.

References

- [1] L. Alfredson, C.-L. Spetz and T. Theorell (1985). Type of occupation and near-future hospitalization for myocardial infarction and some other diagnosis. *International Journal of Epidemiology* 14 (3), 378–388.
- [2] Official Statistics of Sweden (1985). The working environment in figures. Yearbook of Environmental Statistics. Stockholm: Statistics Sweden.
- [3] Waddell, V. and C.D.'A.J. Holman (1985). A population-based study of hospital morbidity by occupation in western Australian males ages 15–64 Years 1981–82. Perth. Health Department of Western Australia.
- [4] Christau, B., M.E. Brammer, B. de Renoud and B. Tardini (1987). Sygdomsforløb-grundlaget for fremtidens sygehusstatistik. *Ugeskr læger* 149/50, 378–388.
- [5] Sundhedsstyrelsen, Dansk Institut for Klinisk Epidemiologi, Epidemiologisk Forskningsgruppe (1984). Evaluering af Landspatientregistret. En pilotundersøgelse. Sundhedsstyrelsen, København. Sygehusstatistik II, 19.
- [6] Buus, H. (1986). Gummi-vulkanisering og sundhedsskader. København: Arbejds miljøfondet.
- [7] Nordic Statistical Secretariat (1988). Occupational mortality in the Nordic Countries 1971–1980. Copenhagen: Statistical Reports of the Nordic Countries.
- [8] Andersen, O. (1985). Dødelighed og erhverv 1970–80. Copenhagen: Danmarks Statistik.