Reliability of Record Linkage in the Swedish Cancer-Environment Register

K. Wiklund & G. Eklund

To cite this article: K. Wiklund & G. Eklund (1986) Reliability of Record Linkage in the Swedish Cancer-Environment Register, Acta Radiologica: Oncology, 25:1, 11-14, DOI: 10.3109/02841868609136369

To link to this article: https://doi.org/10.3109/02841868609136369

Published online: 08 Jul 2009.
RELIABILITY OF RECORD LINKAGE IN THE SWEDISH CANCER-ENVIRONMENT REGISTER

K. WIKLUND and G. EKLUND

Abstract

The Swedish Cancer-Environment Register (CER) is intended to be used for studies on occupational cancer. CER was established by a computerized record linkage between the Swedish National Cancer Register (SCR) for 1961–1973 and the 1960 Population and Housing Census (PH 60) in order to obtain information on occupation, occupational status, economic activity (industry), place of work, domicile, etc. that might be useful for epidemiologic studies on cancer. Accurate inter-register linkage at individual level necessitates high quality of identification in both registers. The hit completeness in this respect was evaluated as 98.8 per cent, a figure regarded as acceptable for most studies based on the CER. Since a hit between SCR and PH 60 did not guarantee accuracy, the reliability of the record linkage was studied in random sample. With the aid of local and national population registries, 0.45 per cent (95% confidence interval 0.21-0.86%) false linkages were revealed in the sample. The CER was in 1982 supplemented with data notified to the SCR 1974–1979, and the hit completeness was 99.2 per cent for the total period. There are good grounds for assuming a lower rate of false linkages for the later period.

Key words: Neoplasms, malignant; registers, occupation, record linkage.

In the early 1970s, attention was directed in Sweden to the feasibility of epidemiologic register studies at national level in regard to possible associations between cancer and occupation but also domicile. The quality of data in the Swedish Cancer Register (SCR) concerning occupation and domicile, however, was judged to be inadequate for such studies. Data concerning occupation was often lacking and when stated it was usually present occupation. Since cancer is assumed to have a long latency period, often several decades, occupational data earlier in life would be more suitable. Domicile at parish-level was registered only in the 1970s.

To improve and supplement the relevant information in the SCR, the 1960 Population and Housing Census (PH 60) was selected, since it contained detailed information on occupation and domicile and other data of potential interest for epidemiologic studies on causes of cancer. Of great importance in this decision was the fact that PH 60 was the first census stored in an electronic retrieval system.

For the period 1961–1973 there were almost 386,000 records in the SCR that could be supplemented with data from PH 60. Such magnitude of supplementation was possible only with computerized record linkage.

Accurate inter-register linkage of personal data requires high quality of individual identification in both ‘parent’ registers. In this respect the unique civic registration number given to all persons in Sweden at birth or immigration was of pivotal importance. This ten-digit number is composed of the year, month and day of birth, three digits coding place of birth (or domicile in 1947 for those born before that year) and sex, and final check digit.

Many Swedish morbidity registers were defective as regards identification data during the 1960s, something which could cause difficulties e.g. in follow-up of occupational cohorts. These difficulties, however, were smaller in Sweden than in the USA, Canada, Great Britain, and several other countries (1, 2, 7–10). In Sweden, these defects were reduced considerably during the 1970s. In the SCR, the importance of tracing cases resulted in great care being taken to improve the quality of the registered civic registration number. This improvement applied both to new notifications and to persons already registered. The current routines at the SCR requires only civic registration numbers for matching against the register.

The Swedish Cancer-Environment Register (CER) was established by computerized record linkage between the SCR and PH 60. Access to the CER is primarily granted

Accepted for publication 16 November 1985.
to researchers studying possible relationships between occupation and cancer. When utilizing information extracted from the CER it is essential to have knowledge of and make reservations for possible sources of error. Documentation of the reliability of the data in the CER is therefore important. The present report concerns results from the computerized record linkage and false linkages in the CER. Some of the sources of error in the two ‘parent’ registers were previously reported (3–6).

Record linkage

Computerized record linkage between the SCR and PH 60 was based on the civic registration numbers. The accuracy and completeness of these numbers in the SCR were checked in a preliminary computerized record linkage, which revealed that approximately 7 per cent of the records in the SCR in 1961–1973 could not be identified in PH 60. The main reasons were incomplete or inaccurate civic registration numbers in the SCR, or change of that number or immigration to Sweden after 1960 (the year of PH 60). Updating of such civic registration numbers was done with the help of local or national population registers, the National Tax Board’s register of immigrants, and other sources of information. In the final merging operation, 3041 of the 385802 records in the SCR for the period 1961–1973 still had incomplete civic registration number and consequently could not be traced in PH 60. The merging additionally revealed 1657 unmerged records which could have been due to incorrect identification number in one or both of the ‘parent’ registries (Table 1). The distribution of these 4698 unmerged records according to site, sex and age has previously been described (6).

The CER was in 1982 supplemented with data notified to the SCR for the period 1974–1979. The figures for the record linkage for the entire period 1961–1979 are also shown in Table 1.

The hit completeness between the SCR and PH 60 was 98.8 per cent for the first period and 99.2 per cent for the whole time, which must be regarded as highly satisfactory. The procedure in establishing the CER was earlier described in detail (11).

False linkages

A hit between a record in the SCR and one in PH 60 did not guarantee accuracy. A hit can be inaccurate if the civic registration number is incorrect in one or both of the registers. A quality check of hits was therefore made, using a random systematic sample.

The selection was made as follows. Records in the SCR were before the record linkage arranged in order of civic registration number. The first record for the sample was chosen from the hits with the aid of a table of random digits, and records were thereafter extracted at intervals of 189 numbers. The planned CER sample size was 2000 persons and the actual number was 1984.

The quality of personal identification in the CER sample was judged by checking the original questionnaire replies for PH 60 against the cancer notification forms. The CER sample was arranged in the same order as the PH 60 forms. The data utilized in the search for the PH 60 forms included domicile (county, municipality, parish, district), location in the population register, year of birth and sex (the PH 60 forms did not include the full civic registration number).

A flow-chart showing checks of personal identification is illustrated in the Figure. The main information extracted from the PH 60 forms, was the full name. One of the PH 60 forms could not be found. The surname was checked against the cancer notification. If there was full agreement in this as well as in other data, the hit was judged to be accurate. When surnames of women were discordant, Christian names and marital status were checked. If all the Christian names were the same on the cancer notification and the PH 60 form, and the woman was or had been married (cancer notification data) the hit was classified as correct. Many cancer notifications included also the maiden name, which in certain cases frequently tallied with the surname on the PH 60 form. In some doubtful cases (with only one Christian name or with several common or discordant Christian names, or no declared marital status) the local population registrars (parishes) were consulted. Discordant male names were always checked with population registrars. Such checking was carried through for altogether 10 women and 4 men.

In the sample from the CER, 9 (0.45%) false linkages were revealed. Of the 9 false linkages, 7 were due to inaccurate personal identification in the SCR and the other 2 were due to changed civic registration number after PH 60. The proportion of inaccurate hits in the total CER was calculated on the basis of this figure. Proceeding from the Poisson distribution, the 95 per cent confidence inter-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Merging of the Swedish Cancer Register (SCR) and the 1960 population census (PH 60)</th>
<th>1961–1973</th>
<th>1961–1979</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of persons in the SCR</td>
<td>385 802</td>
<td>621 019</td>
<td></td>
</tr>
<tr>
<td>Thereof born in 1960 or later</td>
<td>1 753</td>
<td>3 176</td>
<td></td>
</tr>
<tr>
<td>Immigrated to Sweden after 1960 or non-traceable at PH 60</td>
<td>3 336</td>
<td>8 986</td>
<td></td>
</tr>
<tr>
<td>The Cancer-Environment Register should therefore contain</td>
<td>380 713</td>
<td>608 857</td>
<td></td>
</tr>
<tr>
<td>Mismatches against PH 60</td>
<td>4 698</td>
<td>4 754</td>
<td></td>
</tr>
<tr>
<td>Thereof incomplete civic registration number</td>
<td>3 041</td>
<td>2 909</td>
<td></td>
</tr>
<tr>
<td>Included in the Cancer-Environment Register</td>
<td>376 015</td>
<td>604 103</td>
<td></td>
</tr>
</tbody>
</table>
The Swedish Cancer-Environment Register

Random sample (1,984)
Surname on PH 60 form checked against cancer notification (1,983)
PH 60 form lost (1)
Discordance (66)
Concordance (1,917)

Women: Checks of Christian name and marital status (62)
Men: Checks with parish records (4)
Concordance (52)
Discordance (6)
Doubtful (4)
Checks with parish records (10)

Flow chart illustrating checks of personal identification in the Cancer-Environment Register. Figures in parentheses denote numbers of individuals.

Table 2
The 9 discordant cases in the sample

<table>
<thead>
<tr>
<th>Year of birth</th>
<th>Year of diagnosis</th>
<th>County code</th>
<th>Sex</th>
<th>Tumour site (ICD, 7th rev., WHO code*)</th>
<th>Histology code (WHO code**)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1876</td>
<td>1964</td>
<td>05</td>
<td>M</td>
<td>181.0</td>
<td>116</td>
</tr>
<tr>
<td>1887</td>
<td>1964</td>
<td>03</td>
<td>F</td>
<td>170</td>
<td>096</td>
</tr>
<tr>
<td>1890</td>
<td>1964</td>
<td>02</td>
<td>M</td>
<td>181.0</td>
<td>146</td>
</tr>
<tr>
<td>1895</td>
<td>1964</td>
<td>0180</td>
<td>F</td>
<td>175.0</td>
<td>096</td>
</tr>
<tr>
<td>1896</td>
<td>1967</td>
<td>21</td>
<td>F</td>
<td>153.3</td>
<td>096</td>
</tr>
<tr>
<td>1900</td>
<td>1971</td>
<td>0180</td>
<td>F</td>
<td>204.0</td>
<td>206</td>
</tr>
<tr>
<td>1913</td>
<td>1970</td>
<td>05</td>
<td>M</td>
<td>151.0</td>
<td>096</td>
</tr>
<tr>
<td>1915</td>
<td>1972</td>
<td>12</td>
<td>F</td>
<td>172</td>
<td>096</td>
</tr>
<tr>
<td>1917</td>
<td>1970</td>
<td>20</td>
<td>M</td>
<td>162.1</td>
<td>146</td>
</tr>
</tbody>
</table>


Discussion
The CER is an important data source for epidemiologic studies on occupational cancer. Among the CER's advantages are that it includes a large number of persons, covers a long time period, is population-based with well identifiable individuals, and permits rapid computer processing which is not expensive in comparison with case control studies. One of the disadvantages is that for fuller and more reliable data on exposure the search must be extended to other sources. The occupational information in the CER concerns a single point in time, viz. the 1960 census, and provides no data on individual occupational histories as such. Persons in the same occupational category may have been exposed to various noxious agents to a greatly varying extent.

The CER contains information provided by clinicians and pathologists in cancer notifications to the National Board of Health and Welfare, together with data supplied by the country's population in the 1960 census questionnaire regarding occupation, economic activity, etc. Such information inevitably has deficiencies, and from both 'parent' registers the CER has inherited miscodings and various inadequacies.

Quantitative evaluation of all data in such a register is not feasible. The aim should nevertheless be to declare the quality of as many as possible of the main items, thereby assisting users of a register in assessing the magnitude of various recognized errors and sources of error.

Statements of quality are particularly important when, as in the present instance, record linkage has been performed with registers designed for other purposes and maintained by different authorities. Such statements reduce the risk that the data will be used in a misleading fashion.

In the present investigation we were concerned with the personal identification data. When the SCR for 1961–1973 was merged with PH 60, 1.2 per cent of the persons in the SCR could not be traced in PH 60. The proportion of false linkages due to defective personal identification data was calculated to 0.45 per cent, and the corresponding 95 per cent confidence interval was 0.21–0.86 per cent. The quality of personal identification data in the SCR improved since 1973, and there are good grounds for assuming an inaccuracy rate in the CER below 0.45 per cent for 1974–1979. This must be regarded as indicating a high degree of reliability.

The results presented here indicate that the small proportion of false linkages and unmerged records lacks major significance for most studies based on the CER.

References