Pergamon

European Journal of Cancer Vol. 33, Suppl. 3, pp. S37–S43, 1997
© 1997 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0959-8049/97 $17.00+0.00

PII: S0959-8049(97)00022-1

Original Paper

Record Linkage in the National Dose Registry of Canada

M.E. Fair

Occupational and Environmental Health Research Section, Statistics Canada, R.H. Coats Building, Stn. 18 R, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

INTRODUCTION

KNOWLEDGE OF the health effects of radiation has been enhanced by several cohort and case-control epidemiological studies that have been carried out in various exposed groups. Several of these studies have involved populations in Canada associated with the nuclear industry. Recently, a cohort of approximately 267 000 individuals in the 1951–1983 National Dose Registry (NDR) of Canada has been assembled using existing historical files. The NDR cohort has been linked to the Canadian Mortality Data Base (CMDB), which is a centralised file representing all deaths occurring in Canada from 1950 to the present. This paper will concentrate on the essential linkage steps required to organise some of the historic data files in a form suitable for this cohort mortality study: more details may be found in a separate report [1].

Public demands to know what delayed risks are associated with various working and living environments has led private industry, the health and research communities, labour unions, and government to intensify their efforts to obtain quantitative human data. In particular, there is a need to study effects of lower doses and dose rates typical of occupational exposures to radiation, and to address public concerns over effects of low doses of radiation in the general population resulting from increased emphasis on nuclear power generation and the growing application of radiation in industry, medicine and research.

In Canada, several mortality investigations have been carried out over the last twenty years by a number of agencies (e.g. Department of National Health and Welfare, Atomic Energy Control Board, National Cancer Institute of Canada) in collaboration with Statistics Canada [2]. These studies have aided in the development of the necessary files, facilities and record linkage methods and software necessary for follow-up studies. Specific examples of the study cohorts being examined include: nuclear power workers and persons involved in research and development in the nuclear power industry, such as employees of Atomic Energy of Canada Limited [3], Ontario uranium and non-uranium miners [4, 5], Newfoundland fluorspar miners [6], miners and refinery workers of Eldorado Nuclear Resources Limited [7], persons involved in the 1953 and 1958 clean-up operations at Chalk River, and personnel who witnessed nuclear tests in Australia and Nevada [8]. Medical exposure is recognised to contribute a substantial proportion of the total radiation received by the population. Radiation-induced breast cancer is of particular concern because of the increased use of mammography. Women aged 35–39 who participated in an individually randomised controlled trial are being followed up to assess the efficacy of the combination of annual screening with mammography, physical examination of the breasts and the teaching of breast self-examination in reducing the rates of death from breast cancer among women [9, 10]. The Canadian fluoroscopy study is a cohort study of tuberculosis patients first treated in Canadian institutions between 1930 and 1952. A substantial proportion of the cohort was exposed to various levels of low-LET ionising radiation through the extensive use of fluoroscopy to monitor artificial pneumothorax amongst such patients. A study of mortality and cancer incidence in this group is in progress [11].

THE FILES USED

In 1984, a mortality study based on the National Dose Registry was set up by the Bureau of Radiation and Medical Devices (BRMD) of the Department of National Health and Welfare, in conjunction with Statistics Canada and the Atomic Energy Control Board. It involved linkage of the 1951–83 National Dose Registry with the 1951–87 Canadian Mortality Data Base files to determine the causes of death. This large cohort of workers are occupationally exposed to radiation. The purpose of the study is to examine the causes of mortality, with specific emphasis on the relationship between certain types of cancer and cumulative radiation exposure.

Since the amount of identifying information available in machine-readable form in the National Dose Registry was sometimes limited, the Social Insurance Number (SIN) master index file was accessed to help supply additional information to uniquely identify individuals. However, this file is available only from 1964 onward.

The identifiers on the National Dose Registry were not always edited, and thus some basic checks were required to ensure the logical sense of variables. There was no unique

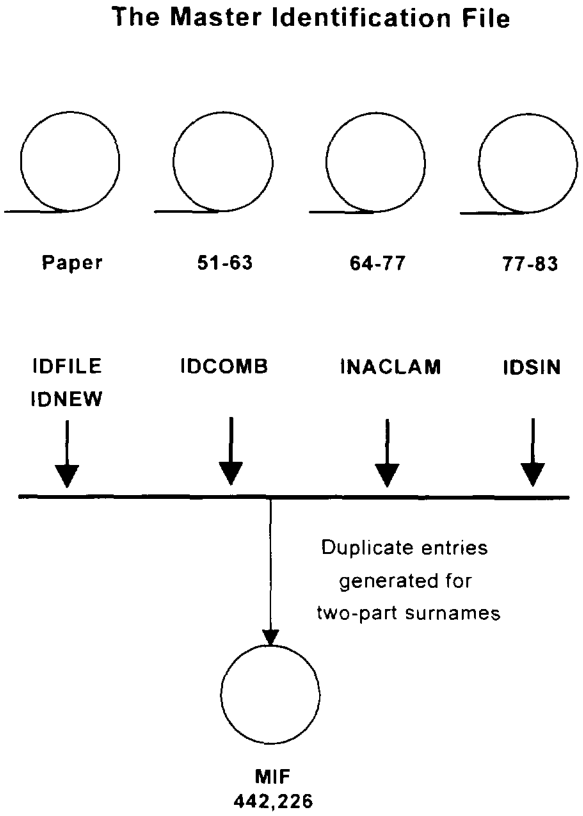
The Master Identification File
Paper
51-63
64-77
77-83
IDFILE
IDNEW
IDCOMB
INACLAM
IDSIN
Duplicate entries generated for two-part surnames
MIF
442,226

Figure 1. Creation of the Master Identification File from the different data sources.

personal lifetime personal identifier available on the records. Over the period, several numbering systems had been utilised to record the annual doses for individuals. There were five different identification files (known as IDFILE, IDNEW, IDCOMB, IDACLAM and IDSIN), which were merged to create a Master Identification File (MIF) of this study (Figure 1).

Concurrent with this work was the development of a new computer system for the National Dose Registry [12]. There is also constant updating of this registry file, which can complicate the generation of appropriate work and exposure histories. The Canadian Mortality Data Base contains records for all registered death events from the provinces and territories since 1950, and is under the custody of Canada's central statistical agency. Statistics Canada (Figure 2). These records exist in both machine-readable and microfiche formats. The machine-readable records consist of provincial records that have been converted into a standardised form [13].

The files on the CMDB used in this study are sequenced by : (i) a phonetic code of the surname (referred to as the NYSIIS code); and (ii) the gender of the deceased. The total number of death registrations from 1951 to 1987 is 5.9 million. The file contains identifying information, plus the date, place and coded underlying cause of death, in effect at the time of death. Some of the personal identifiers on historical death records were missing. The availability of identifiers can vary by province and by year [14].

All studies, such as the National Dose Registry Mortality study, involving long-term medical follow-up carried out at Statistics Canada must satisfy a rigorous review and approval process. The Statistics Act protects the confidentiality of all records.

AN OVERVIEW OF RECORD LINKAGE

Record linkage is the process of bringing together or 'grouping' two or more separately recorded pieces of information pertaining to the same individual or entity. The procedures for a computerised probabilistic record and the generalised record linkage system (GRLS) have been described in detail elsewhere [15–20]. The generalised system that has been developed at Statistics Canada is particularly flexible and permits the introduction of numerous refinements, such as carrying out one-file (e.g. to create individual histories for a person within the National Dose Registry regardless of the company where an individual has worked) and two-file linkages (e.g. linking the dose registry with mortality records).

Often the decision to 'link' particular records depends on the similarities of names, birth dates, birth places and such. Each of these identifiers carries only a limited discriminating power and is fallible, so it may sometimes seem dissimilar when in fact the same person is involved. First the files have to be searched to bring pairs of records together for comparison. Conceptually, each record on file A (e.g. the National Dose Registry) is compared with each record on file B (e.g. the Canadian Mortality Data Base) to form a set of records C, which one attempts to classify as links or non-links. In practice, the files are blocked using identifiers (e.g. by the phonetic code of surname and gender code) to limit the number of pairs compared. A decision is made as to whether the link is true, and then a group of the appropriate records relating to the same individual is formed.

The generalised system estimates how likely it is that a pair of records refer to the same individual. It does this by comparing corresponding fields one at a time between records, and seeing if the values agree, partially agree, disagree or are missing. The type of comparison outcomes can be similar to those that a human clerk would do in carrying out the same task.

Quite briefly, when comparing values Ax from a record A (e.g. an NDR record, which is used to initiate the search) with value By from a record B (e.g. a death record, which is the file being searched), the ODDS in favour of LINK associated with the outcome $Ax{\cdot}By$ (i.e. the comparison pair of values) may be written in terms of the relative probability of occurrence of the particular outcome in LINKS as compared with NONLINKS, that is:

$$\text{ODDS} = \frac{P\,(Ax \cdot By|\text{LINK})}{P\,(Ax \cdot By|\text{NONLINK})}.$$

As in information theory, the odds are usually expressed as logarithms to the base 2, and are often multiplied by ten and rounded to avoid decimals:

$$\text{Outcome Weight} = 10 * \log_2\,(\text{ODDS}).$$

A rule is created to compare the fields (x, y) in the records. The comparisons can be straight comparisons, cross comparisons (e.g. comparing first forename on record x with the second forename on record y), or specially written functions.

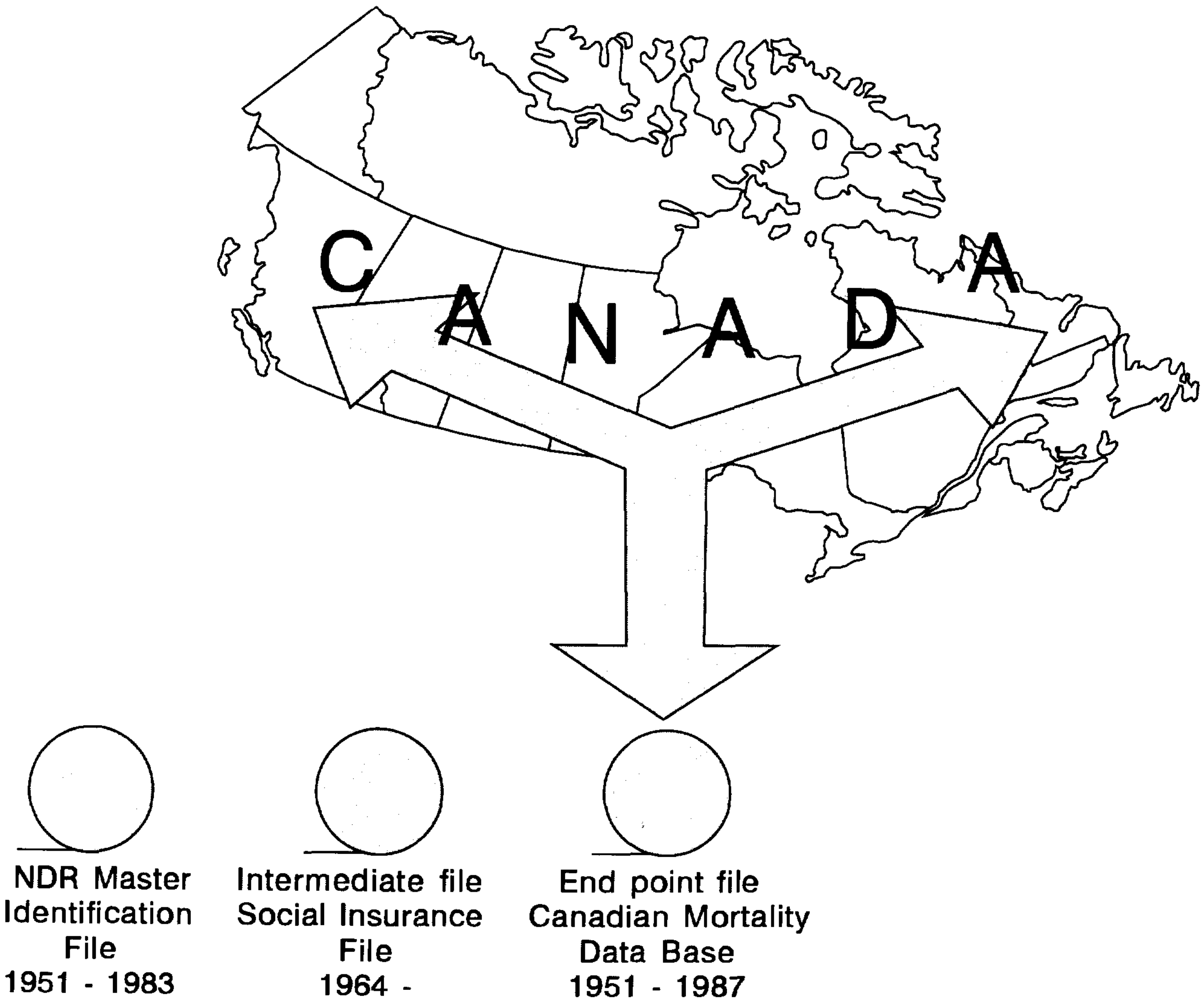
CANADA
NDR Master Identification File 1951 - 1983
Intermediate file Social Insurance File 1964 -
End point file Canadian Mortality Data Base 1951 - 1987

Figure 2. Linkage of the Master Identification File with the Canadian Mortality Data Base.

The total odds in favour of a match may be expressed as the sum of a number of 'outcome weights':

$$\begin{aligned}\text{Total Weight} = {} & \text{Outcome Weight } (O_1) \\ & + \text{Outcome Weight } (O_2)\cdots \\ & + \text{Outcome Weight } (O_n)\end{aligned}$$

where O_1, $O_2 \ldots, O_n$ are the outcomes for the rules 1 to n (including any used for blocking) used to compare fields on the records.

The outcomes are assumed to be statistically independent.

The total weight becomes the overall estimate of how 'probable' it is that the potential link is in fact a definite link. By comparing the total weight against two thresholds this estimate is converted into a decision as to whether or not the link is a 'true' one. If the total weight is above the upper threshold, the link is assigned a temporary status of 'definite link'; if it is below the lower threshold, the temporary status is 'unlinked'; if it is between the two thresholds the temporary status is 'possible':

```
<········· (Lower ········· Upper) ·········>
 Unlinked          Possibly          Linked
                    Linked
```

Possible links are then examined in more detail, perhaps on a sample basis, to fine-tune the setting of the thresholds. In smaller projects, manual resolution can be carried out on these links. Further reference is usually made to the original death registration source documents where further identifying information may be available, or by reference to other source documents relating to the cohort file. In large projects, one threshold value is sometimes chosen to classify records as links and non-links.

The selection of thresholds involves two types of error. A Type I error occurs if an unlinked pair is erroneously classified as linked because it falls above the upper threshold (i.e. false positives). A Type II error occurs if a pair is erroneously classified as non-linked (i.e. false negatives). The upper and lower thresholds are determined by the setting of acceptable error bounds to limit the number of errors for the analysis, as is determined to be appropriate.

Where linkages can be based on some sort of personal identifying number, there is much greater certainty of achieving a correct match. However, that number may be incorrectly recorded on some records so that disagreement of it is not necessarily proof of a non-match. Such numbers are occasionally improperly 'borrowed' from other people so that agreement is not always positive proof of a correct match. Although rare, such occurrences make it prudent to check names and other usual identifiers as a backup when doing numerical linkages. Thus, the method of probabilistic record linkage should still be applied to improve the accuracy of even the numerical linkages.

THE VARIOUS LINKAGE STEPS IN THE NDR STUDY

For the National Dose Registry Mortality study, there were several types of linkages required to bring together the appropriate personal identifiers, dose histories, and death information.

Internal linkage—creating the cohort file

The Master Identification File (referred to as the 'identification' file) contained more records than there were people. Duplicate records required consolidation with appropriate provisions for cross-referencing entries where there have been changes of name and/or where key identifiers have been reported differently on separate occasions. After a series of internal linkages (i.e. by numerical identification, and then later by personal identification), the result was the cohort of individuals to be included in the study.

Social Insurance Number master index linkage

Even after the series of internal linkages, the availability of personal identifiers was still relatively poor for parts of the cohort. These identifiers are vital for a successful death linkage. Permission was granted for Statistics Canada to use the Social Insurance Number master index for the sole purpose of unambiguously identifying cohort members.

The identification file had to be matched to the index file, for records having a social insurance number, in order to extract the required personal identifying variables. The linkage was conducted to confirm that the matched individual was indeed the correct individual. Data abstracted from the index file were strictly for internal use only.

Dose history linkage

Since the cohort file was originally sent to Statistics Canada, the Department of National Health and Welfare has been conducting dynamic merges to their database. This will help to regroup dose records, reduce the number of fragmented records, and consolidate records into comprehensive dose histories for each study member. The file resulting from the internal linkages indicated which records on the NDR appear to belong to the same individual. A file containing original master identification file information (surname, and numerical identifiers) and the Statistics Canada sequence number was sent to the Department of Health and Welfare to allow the creation of the dose history file required later in the study.

Death linkage

The internally linked cohort was matched against the mortality records. By linking the two files, it is possible to measure the cohort members subsequent risk of death, and to add information regarding the fact, underlying cause, place of death, place of birth, and birth year information.

Preparation of the analysis file

A match of data from the identification, mortality and dose history files was performed to create a comprehensive record for each member of the study cohort. Where the information was available, each record included month and year of birth, gender, the death data, an availability word, the death linkage weight, and a dose history. The availability word indicates whether items were stated or blank on the records being compared. Any unmatched records from the identification or dose history files underwent special scrutiny.

RESULTS AND RECOMMENDATIONS ARISING FROM THE STUDY TO DATE

The record linkage work for the National Dose Registry Study has been divided into seven stages as shown in Figure 3. The details of this work have been described elsewhere [1]. The results of each stage will be described, and finally some recommendations will be made about optimal designs of centralised dose registries for epidemiological follow-up purposes.

Stage I: Evaluation of the NDR Master Identification File

Five files of identifying particulars and other data that would be helpful during the mortality linkage were provided to Statistics Canada by the Department of National Health and Welfare (see Figure 3). Duplicate entries were created for records having two-part surnames, bringing the total number of records in the identification file to 442,226.

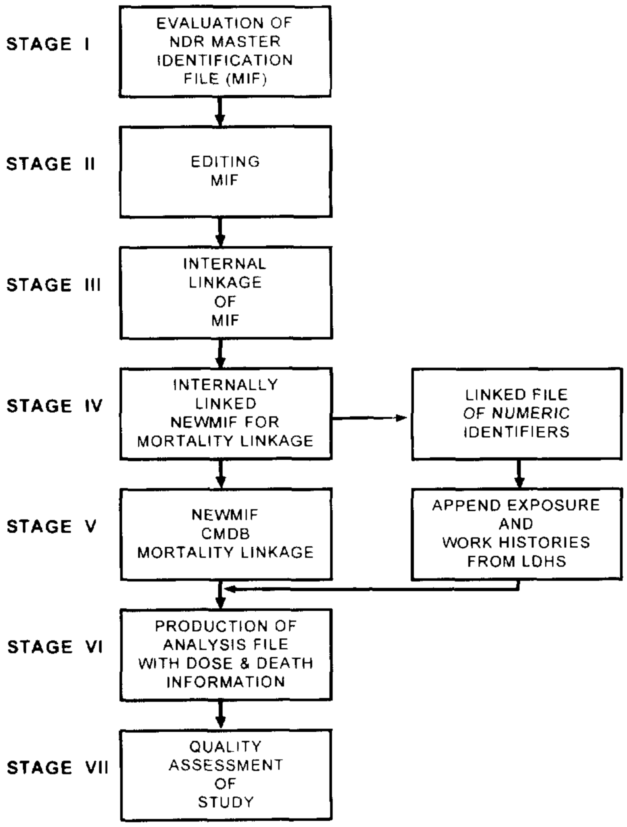
STAGE I
EVALUATION OF NDR MASTER IDENTIFICATION FILE (MIF)
STAGE II
EDITING MIF
STAGE III
INTERNAL LINKAGE OF MIF
STAGE IV
INTERNALLY LINKED NEWMIF FOR MORTALITY LINKAGE
LINKED FILE OF NUMERIC IDENTIFIERS
STAGE V
NEWMIF CMDB MORTALITY LINKAGE
APPEND EXPOSURE AND WORK HISTORIES FROM LDHS
STAGE VI
PRODUCTION OF ANALYSIS FILE WITH DOSE & DEATH INFORMATION
STAGE VII
QUALITY ASSESSMENT OF STUDY

Figure 3. The seven stages of the NDR mortality study.

Checks were made to detect inconsistencies in coding and data entry. For example, some names contained characters other than A–Z, and some months had values greater than 12. Initially, 12,000 records were returned for correction, particularly to do with variables essential for the mortality linkage (e.g. surname, given name, gender, and complete birth date).

Stage II: Editing and correcting the Master Identification File

Customised computer programs were written to check discrepancies in gender coding and in given names. A dictionary was created which gives the frequency of occurrence of given names by gender. The probability of the forenames relating to a male or female are calculated, and a gender code is derived based on the highest odds. This gender code is then compared to that available on the record. Where the gender code had been derived, but disagreed, these were then reviewed and corrected if necessary. Given names were also edited.

Stage III: Internal linkage in the Master Identification File

One of the major tasks was to create a file of individuals that defined a cohort of special interest and to create a corresponding file of personal employment and exposure histories. Records were fragmented in the Master Identification File for two main reasons: (1) there were multiple entries for persons who had left and rejoined the workforce at different times; and (2) name changes, especially for married women and nuns. The purpose of the internal linkage was to assign a common number to all records relating to the same individual.

Several internal linkages were carried out. A numeric linkage was the first, as these are inexpensive and easy to perform. The identification record could contain one or more of the following numerical identifiers: Namecode, Radiation Protection Bureau number or Social Insurance Number. The second internal linkage used the phonetic code of the surname [21] plus the gender code as the combined pocket identifier within which individuals were searched. At this time, significant additional data became available to the NDR. Atomic Energy of Canada Limited (AECL) did not send its dose records to the National Dose Registry on a yearly basis until 1981. A file containing pre-1981 records from AECL was given to the NDR to link to the existing identification file. Subsequent to the two numeric linkages above, 18,057 records from Atomic Energy of Canada Limited were added to the identification file. These records were matched to the existing identification file by SIN where available. Many of the records in the file were virtually the same, and it was decided to combine records having no conflicting identification into a single record so that superfluous records could be deleted. This process left approximately 437,599 records corresponding to a possible 371,849 individuals. This collapsed identification file was renamed by Statistics Canada as the new master identification file, NEWMIF, to avoid confusing it with the original identification file.

Collapsing the groups was not sufficient to increase the availability of data to a level necessary for linkage to the deaths. To improve the availability of identifiers, it was necessary to match records having a valid SIN with the SIN Master Index File. When an exact match (SIN, surname, etc.) was found, information was shared between the records, creating hybrids. Duplicate entries were deleted leaving a total of 534,159 records (235,166 unchanged NDR records, 203,855 hybrids, and 95,138 SIN records). From the NEWMIF, 64,007 groups (individuals) were split off because they did not meet the minimum data required (surname, first initial and birth year) to participate in the mortality linkage. A tape file of these records was sent to Health and Welfare for further examination. A total of 31,054 groups were dysfunctional in that they had no numerical keys or no dose records. The remaining 32,953 groups that apparently linked to dose records may have already existed on the NEWMIF, or may have more identifying data available to permit their re-entry into the study at some point. The answers to these concerns are pending. At present, they will not affect the number of individuals in the current cohort, ready for the mortality linkage.

Stage IV: Merging of NEWMIF with exposures and work histories

After the final internal linkage by a phonetic surname code (NYSIIS), and all updates were finalised, the NEWMIF to be used for the mortality linkage contained 464,974 records representing an estimated cohort of 266,465 individuals. This number does not include those 54,007 individuals split off earlier. This number of individuals will also vary slightly after further checks with corresponding dose histories are made.

A file containing the group sequence number, surname and the original numerical identifiers was sent to the Department of Health and Welfare for two purposes: (i) to add any updated information on birth data, province of last contact, and year of last contact, resulting from internal maintenance of the NDR; and (ii) to identify the cohort members for which dose histories will be required. Statistics Canada appended this updated information in (i) to the corresponding records on the NEWMIF. This recent information from the updated National Dose Registry was particularly helpful in the mortality search.

Stage V: Linkage of the new identification file (NEWMIF) with the death file (CMDB)

A file of about 465,000 NEWMIF records was matched with about 5.7 million death records. Since large files were involved, this linkage was split by gender code, so that males and females could be linked separately to the CMDB. All potentially linkable records were extracted. A lower threshold (−30), below which there was unlikely to be any correct links, was established. Samples were drawn to aid in the establishment of suitable thresholds, depending on the availability of linkage identifiers.

Phase VI: Production of the analysis file with death, dose and work histories

An analysis file was created containing the required information for each individual. All unique numbers and names were deleted from the file. A new number was assigned for each record and its corresponding dose record to facilitate checking, should a need arise. Some anomalies are currently being investigated, where the dose and work histories did not correspondingly merge together.

Phase VII: Quality assessment of the study

Further analysis may be carried out to explore, measure, and document the ways in which thresholds were selected and the linkage errors occurring at various stages. There are a number of approaches that might be used. For example, manual resolution, with reference to source documents that may have additional information, may assist in the estimation of error rates in the setting of thresholds. In addition, other approaches might include comparing the methods used here with other results from sub-cohorts within the file that have been linked earlier, or that help confirm the vital status of the individual (e.g. drivers licence, income tax files etc.).

Recommendations

A number of recommendations for a centralised dose registry arise as a result of the experience of the various phases in this study. (a) A centralised dose registry should contain a unique lifetime number for the individual plus surname at birth and currently complete forenames, birth date, birth place and gender code. The birth year and gender code are generally essential in the analysis. (b) The data should be edited, and particular attention should be given to the handling of surnames and forenames in a standardised manner. A typical set of edit rules for items has been outlined elsewhere [18]. (c) All records relating to the same individual should be brought together. The effects that errors in this phase will produce in both the linkage with mortality and in the analysis should not be underrated. For example, if the internal linkage fails to bring together the records of women when the maiden and married names differ, then the number of individuals in the cohort will be inflated, and the mortality search may fail to find a linkage for one of the records. If records are incorrectly combined, then the work and exposure histories will be incorrect, and the mortality search may be less effective. (d) An ongoing centralised registry should phonetically code surnames and have the date and province of last contact available to facilitate the follow-up of individuals. (e) If one wishes to ascertain the causes of death and cancer, it is necessary to have this information recorded in accessible media, preferably centralised and in machine-readable form. (f) The development of individual exposure and work histories are required over time. (g) Sources of error should be understood and measured.

The experiences of cancer registration principles and methods may be helpful in setting up national dose registries [22], particularly as it relates to items of patient identification (Chapter 6).

Overall, the idea underlying the study is quite simple. Records of the National Dose Registry are accessible by computer as are centralised death records. Linkages of the two sorts of records on an individual basis is dependent, in principle, on the availability of sufficient personal identifying information common to the two sorts of records, to ensure against ambiguous matches and mistaken identities. Analysis of cause of death, cumulative individual radiation exposure and work histories is then possible.

1. Coppock E, Dobson D, Fair ME. Occupational radiation exposure and mortality study: I. Internal linkage of the Canadian National Dose Registry. Box 1046, Ottawa, Ontario, K1P 5S9. Atomic Energy Control Board. AECB Publication no. Info-0417-1, 1992.
2. Statistics Canada. Occupational and Environmental Health Research Section—Studies and references relating to the uses of the Canadian Mortality Data Base. Available from: Occupational and Environmental Health Research Section, Stn 18R, R. H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Statistics Canada, 1992.
3. Gribbon MA, Howe GR, Weeks JL. A study of the mortality of AECL employees V. The second analysis: mortality during the period 1950–1985. Pinewa, Manitoba R0E 1L0, Whiteshell Nuclear Research Establishment, Atomic Energy of Canada Limited publication no. AECL-10615, 1992.
4. Muller J, Kusiak R, Ritchie AC. Factors modifying lung cancer risk in Ontario uranium miners. Ontario Ministry of Labour, Ontario Workers' Compensation Board, Atomic Energy Control Board of Canada, Toronto, Ontario, 1989.
5. Kusiak RA, Springer J, Ritchie AC, Muller J. Carcinoma of the lung in Ontario gold miners: possible etiological factors. *Br J Ind Med* 1991, **48**, 808–817.
6. Morrison HI, Semenciw RW, Mao Y, Wigle DT. Cancer mortality among a group of fluorspar miners exposed to radon progeny. *Am J Epid* 1988, **128**, 1266–1274.
7. Newcombe HB, Smith ME, Howe GR, *et al.* Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers Biol Med* 1983, **13**, 157–169.
8. Raman S, Dulberg CS, Spasoff RA. Mortality study of Canadian military personnel exposed to radiation: atomic test blasts and Chalk River nuclear reactor clean-up, 1950s. Ottawa, Department of Epidemiology and Community Medicine, Faculty of Health Sciences, University of Ottawa, 1984.
9. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Can Med Assoc J* 1992, **147**, 1459–1476.
10. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Can Med Assoc J* 1992, **147**, 1477–1488.
11. Howe GR. Breast cancer incidence and mortality in the Canadian fluoroscopy study: the establishment of computerized record linkage facilities for the National Cancer Incidence Reporting System (1975–1983). Box 1046, Ottawa, Ontario, K1P 5S9; Atomic Energy Control Board. AECB publication no. 7. 1221, 1993.
12. Ashmore JP, Krewski D, Zielinski JM. Protocol for a cohort mortality study of occupational radiation exposure based on the National Dose Registry of Canada. *Eur J Cancer* 1997, **33**(Suppl. 3), S10–S21.
13. Smith ME, Newcombe HB. Use of the Canadian Mortality Data Base for epidemiological follow-up. *Can J Public Health* 1982, **73**, 39–46.
14. Lalonde P. Availability tables for the Canadian Mortality Data Base (CMDB), 1950–1989. Available from: Occupational and Environmental Health Research Section, Stn 18R, R. H. Coats Building, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. OEHRS - NO 10, 1991.
15. Smith ME, Silins J. Generalized iterative record linkage system. *Proceedings of the Social Statistics Section*, 1981. American Statistical Association, Washington, D.C., 1981, 128–137.
16. Howe GR, Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. *Comput Biomed Res* 1981, **14**, 327–340.
17. Newcombe HB. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business.* Oxford, Oxford University Press, 1988.
18. Carpenter M, Fair ME. A standard data collection package for medical follow-up studies. Ottawa, Ontario, K1A 0T6, Statistics Canada, Canadian Centre for Health Information. *Health Reports* (Cat. 82-003). 1990, **2**, 157–173.
19. Statistics Canada. Generalized record linking system concepts. Available from General Systems Development Division, R. H. Coats Building, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, 1991.
20. Newcombe HB, Fair ME, Lalonde P. The use of names for linking personal records. *J Am Stat Assoc* 1992, **87**, 1193–1204.

21. Lalonde P, Fair ME, Carpenter M, Scott T. Name encoding schemes. Box 1046, Ottawa, Ontario, K1P 5S9. Atomic Energy Control Board. AECB publication no. INFO-0418, 1992.

22. Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG (eds.). Cancer registration principles and methods. Lyon, France, International Agency for Research on Cancer. IARC Scientific Publication No. 95, 1991.

Acknowledgements—This project has been a long-term undertaking having commenced in 1984. Many people have contributed to the work described in this paper. The author gratefully acknowledges the assistance of the following people from Statistics Canada: C. Poliquin, M Werner, D. Dobson, L. Coppock, G. Quesnel, P. Lalonde; from the Bureau of Radiation and Medical Devices: P. Ashmore, B. Davies. Dr Howard Newcombe served as consultant in the early planning stages of this work and we thank him for his helpful comments on this paper. Funding has been made available from the Atomic Energy Control Board, the Department of National Health and Welfare, and Statistics Canada. The various provincial and territorial vital statistics registrars have given their permission for the use of the mortality data required for the analysis.