Appendix 3 (c) Cancer registration in Ontario: a computer approach

E.A. Clarke, L.D. Marrett and N. Kreiger

Ontario Cancer Registry, Ontario Cancer Treatment and Research Foundation, Toronto, Canada

Background

The Ontario Cancer Registry (OCR) is a population-based registry covering the entire province of Ontario. Ontario is the most populous province in Canada, with 9.1 million people in 1986 (Statistics Canada, 1987) and an area of over one million square kilometres; 82% of the population inhabit the urban areas, mostly in the southern part of the province. Although 80% of the residents were born in Canada, they represent a wide variety of ethnic groups of which the largest are British, French, Italian and German.

The OCR is operated by the Ontario Cancer Treatment and Research Foundation, which was incorporated in 1943 by an Act of the Legislature of the Province of Ontario (The Cancer Act) 'to establish a program of cancer diagnosis, treatment and research' in the province. This act followed a recommendation by a provincial commission that radiotherapy, then the most effective method of cancer treatment other than surgery, be centralized. Regional cancer centres (RCCs) were therefore established in major cities across the province to provide radiotherapy to outpatients. In addition, the Ontario Cancer Institute, incorporating the Princess Margaret Hospital (PMH), was established in Toronto in 1958. Together, the RCCs and the PMH provide all the radiation therapy for cancer patients in the province, as well as chemotherapy and consultative services for approximately 50% of cancer patients in Ontario.

The Ontario Cancer Treatment and Research Foundation, including the OCR, is supported primarily by the Ontario Ministry of Health (MOH). Patient care is publicly financed; in Ontario about 95% of Ontario residents are covered by a comprehensive government health insurance plan. While some residents of Ontario seek medical care outside the province, the proportion of claims for in-patient care originating from outside Ontario is less than 1%. The majority of such claims are made by residents of Ontario who live close to its borders.

The Cancer Act of 1943 included provision for 'the adequate reporting of cancer cases and the recording and compilation of data'. Cancer is not a legally reportable disease in Ontario, but amendments to the Cancer Act since 1943 have provided legal protection for organizations or individuals in the health-care system who report

information on cases of cancer to the Ontario Cancer Treatment and Research Foundation. These amendments enable information in the OCR to be used for epidemiological and medical research. In addition, each hospital in the province is required to forward diagnostic information on every discharged patient to the MOH. The MOH uses this information for administrative purposes and provides the OCR with copies of data on cancer patients; thus, a degree of compulsory reporting is in effect for hospitalized patients.

The process of cancer registration

Although the OCR includes cancer patients diagnosed since 1964, there was a major change in registration methods in 1972. Only registration techniques employed since 1972 will be described in the remainder of this report. Details of methods used in earlier years may be found in a monograph on the first twenty years of Ontario cancer incidence data (Clarke *et al.*, 1987). It should be noted that the OCR does not attempt to register non-melanotic skin cancers.

The OCR is created entirely from records generated for purposes other than cancer registration supplied from a variety of sources. A computerized record linkage system brings together these sources, and multiple records pertaining to the same individual are linked. A set of computerized rules known as the Case Resolution system is then applied to the linked records, which allocates the appropriate site of disease, histology, date and method of diagnosis, residence, and other information for each case of cancer. These methods result from a collaboration between two departments of the Ontario Cancer Treatment and Research Foundation, namely, Epidemiology and Statistics and Information Systems.

Sources of data

Four major sources of data are employed to create the OCR:

- hospital separations with cancer as a diagnosis;
- pathology reports with a mention of cancer;
- death certificates in which cancer was the underlying cause of death;
- reports on patients referred to the RCCs and PMH.

Hospital separation reports

Hospital in-patient separation data with mention of cancer are forwarded to the OCR by the MOH. These were submitted as documents until 1975, after which time the data were provided on magnetic tape. In 1978, the MOH instituted a requirement that each hospital submit an abstract for each discharge to an independent organization, the Hospital Medical Records Institute (HMRI). The HMRI abstract form provides for the recording of sixteen possible discharge diagnoses (as opposed to the single diagnosis permitted on hospital separation forms prior to 1978) but these abstracts do not contain surnames or given names. After processing (which includes some editing), HMRI forwards the resulting file to the MOH where name and Ontario Health Insurance Plan (OHIP) number are added. A subset of this integrated file is created,

consisting of records in which cancer is one of the discharge diagnoses, and this file is forwarded annually to the OCR. Currently, about 100 000 hospital separations are received each year.

Pathology reports

In 1973, pathology laboratories across the province were asked to submit copies of reports in which cancer was mentioned. By 1980 all were complying. The annual number of pathology reports received by the OCR has increased dramatically from less than 15 000 in 1973 to about 50 000 in recent years. Paper records are provided to the OCR by participating laboratories and are coded by OCR staff.

Deaths

The OCR has data in machine-readable form on all deaths of Ontario residents. For the years 1972–80, these data were received from Statistics Canada, by special arrangement with the Office of the Registrar General of Ontario. Since 1981, the Office of the Registrar General of Ontario has annually provided a computer tape directly to the OCR. Underlying cause of death is coded by trained nosologists in the Office of the Registrar General. All deaths with cancer considered to be the underlying cause are included in the OCR. There were about 11 500 cancer deaths in Ontario residents in 1972 and 17 000 in 1986.

Treatment centres

Initially, abstract cards recording minimal information on their cancer patients were completed at each RCC and the PMH. Those from the RCCs were forwarded to the OCR for further data abstraction and coding. Between 1972 and 1981, these cards were gradually discontinued at the PMH and the RCCs, and appropriate data were subsequently forwarded to the OCR in machine-readable form. Abstract cards were also created for tumour registries maintained at the RCCs for cases diagnosed in their regions but not referred to the centres. These cards were forwarded to the OCR for abstracting and coding until the registries were discontinued by the RCCs in 1976. The OCR receives about 20 000 reports on cancer patients from the RCCs and PMH each year.

Coding, data entry and preprocessing of data

All cancer records submitted to the OCR in the early years (1972–1975), except death records in which cancer was reported as the underlying cause, were coded and entered into the computer centrally by the OCR. Between 1975 and 1977, hospital discharge information was coded at the MOH. Since 1978, it has been coded in the medical records departments of hospitals in Ontario. These data have been sent to the OCR on magnetic tape since 1975. Given the fact that a passive system of cancer registration is employed, it is not possible, for the most part, to institute formal methods of quality control with regard to coding.

Pathology reports have always been coded and the data entered by clerks at the OCR. These are subjected to routine assessment of quality, as were other records previously coded at the OCR. Difficult reports are circulated among coding staff and are discussed at regular meetings with the medical staff.

Data from the RCCs and PMH have been collected uniformly since their establishment. With computerization of records at these centres, coding has devolved to their medical record staffs. The managers of health records at each RCC and the PMH meet twice a year to discuss coding and other quality control issues. The RCCs also send copies of pathology reports and a clinical description of the cancer to the OCR. These reports are recoded, and any discrepancies are corrected after discussion between the RCC and the OCR.

Routine quality control of the data entry phase is carried out on all records of the OCR. Samples of reports entered online are verified by routine recoding and key entry. The data entry system requires that certain variables (e.g., surname of patient, date of diagnosis, site of disease) always be entered. As data are entered, they are edited for validity, consistency and plausibility. Data received on magnetic tapes are also subjected to the same editing procedures (edits); however, these are carried out by batch programs. Validity edits reject data which are inherently incorrect (e.g., the 13th month, the 32nd day). Consistency edits compare two or more data fields and report contradictions (e.g., a male patient with ovarian cancer, a treatment date preceding date of birth). Edits for plausibility report unlikely but possible situations which are potential errors (e.g., a 110-year-old patient, a five-year-old male with prostatic cancer). These plausibility edits are checked manually and corrected if necessary. Coded data (e.g., residence, hospital, birthplace) are compared with tables constructed by the OCR specifically for validation purposes. Finally, numerical data are validated with check digits.

Site of cancer on all records has been coded to the Eighth Revision of the International Classification of Diseases (ICD-8) (WHO, 1967) prior to 1979 and to the Ninth Revision (ICD-9) (WHO, 1977) since that time. In addition to the computer edits described above, all ICD codes are converted during processing to ICD-9. Before 1979, morphology was coded to the Manual of Tumor Nomenclature and Coding (MOTNAC) (Percy *et al.*, 1968) and, since 1979, to the International Classification of Diseases for Oncology (ICD-O) (WHO, 1976b). MOTNAC codes are also converted to ICD-O morphology (M) codes by computer.

Linkage

Once the source files have been preprocessed, all records pertaining to an individual are linked together by a sequential computer linkage. In order to link together this large volume of data, the OCR has developed a sophisticated computer record linkage system based on the Generalized Iterative Record Linkage System (GIRLS) designed by Statistics Canada in conjunction with the Epidemiology Unit of the National Cancer Institute of Canada (Howe & Lindsay, 1981).

Since Ontario does not have a unique number in the health or political system which identifies an individual throughout life, linkage is based on a number of

identifying variables including name, date of birth, OHIP number, hospital where diagnosed and hospital chart number. It should be noted that an OHIP number is allocated to a family, and does not distinguish between individual members of that family. When a child reaches the age of 18 (or 21, if attending university), he or she is assigned a new OHIP number. Change of employment, or divorce, may also result in the allocation of new OHIP numbers to individuals.

The present computer linkage is completed in several stages. First, a New York State Identification Intelligence System (NYSIIS) code is created, which is a phonetic version of the surname. Only records which have the same NYSIIS code are compared for possible linkage; therefore, records with names having similar spellings but different NYSIIS codes do not have an opportunity to link. Records with the same NYSIIS code constitute a pocket within which records are compared. A numerical score or weight is assigned to each variable when two records are compared. The greater the sum of the weights of the variables compared, the greater the probability that two records linked by the system belong to the same individual. The word iterative in the acronym GIRLS indicates that this process of allocating weights is repeated more than once. The system uses previous observations to assign more precise weights.

Each link (i.e., each pair of records brought together) is classified into one of three categories: definite, possible or rejected, based on the magnitude of the total weight. The distribution of the total weights in the linked file is usually bimodal, clustering around a high weight (definite, i.e., likely to be true links) and a low weight (rejected, i.e., unlikely to be true links). The middle range of weights contains possible links, i.e., those in which it is uncertain that paired records relate to the same individual. Linked records in this range (the grey area) are reviewed by health record staff of the OCR who have access to additional data that were not used in the linkage. An example would be information contained in the complete pathology report which might confirm the suspicion by the staff that an earlier biopsy had been performed. Decisions are made to accept or to reject each link in the grey area and the result is then entered into the linked files. This manual resolution reduces the number of false links accepted and missed true links, but both still occur. The size of the grey area varies according to the files being linked; 2–12% of potential links are manually resolved.

Linkages of source files are performed in sequence (see Figure 1). Each year, hospital reports are linked internally to bring together multiple admissions for the same patient. Pathology reports are then linked to these aggregated hospital records, since most pathology reports will be related to a hospital stay. This combined hospital-pathology file is subsequently linked with previous years' incidence to identify incident (as opposed to prevalent) cases, producing provisional incidence data. Every second or third year, deaths due to cancer and records from the RCCs and the PMH are linked to these provisional data. These final linkages add few new cases of cancer, although RCC and PMH records improve data quality, particularly the specificity of site and histology.

Finally an internal linkage is performed on the entire file using pockets other than those created by NYSIIS codes. This linkage allows groups of records with different

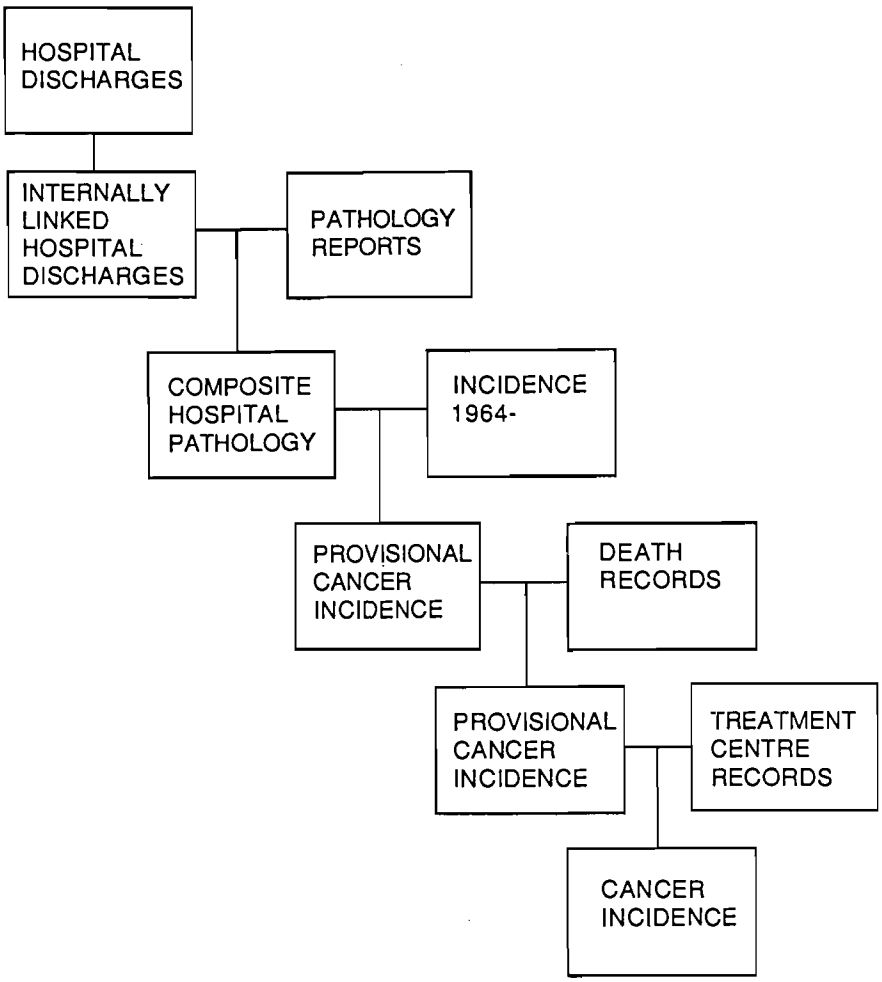

Figure 1. Sequence of linkage of source files

NYSIIS codes to be compared so that records which may pertain to the same patient have the opportunity to link. There are three distinct phases to this linkage. Within pockets created in each phase, comparisons of all possible pairs of records are carried out and weights are assigned, as in other linkages. In the first phase, pockets are assigned using OHIP number and sex. In the second phase, pockets are formed using birth year and the first three letters of the given name. The third phase utilizes the first three characters of the surname. Records linking at a high weight in one phase are not included in subsequent phases. The grey area resulting from this three-phase linkage is resolved as in other routine linkages. These linkages reduce the effect on the OCR of errors in spelling or in transcription of surnames.

Nearly all cancer patients have multiple source records. A set of computer programs has been developed by OCR staff to create a composite identification record containing the best identifying information from all source records on a patient (e.g., surname, given names, date of birth, sex). This is then carried forward into the next linkage. These programs also find conflicts between individual source records that may be the result of false links which had not been identified earlier. These conflicts are reported and reviewed by OCR health record staff, who make corrections as indicated.

Allocation of site, histology and other information

Groups of linked source records for individual patients are processed by a second major system, Case Resolution. This consists of a series of computer modules developed by OCR staff, and applies medical logic to the source records for a patient to determine the appropriate site of disease, histology and date of diagnosis, since these may vary between source records.

In the Case Resolution system, cancer sites on all source records belonging to one individual are examined to determine the most specific site in a rubric or group of rubrics of the ICD. Only the most specific site codes are retained for further processing. Thus, if one source record indicated 'malignant neoplasm of digestive tract' (ICD-9 159.9), another indicated 'malignant neoplasm of stomach not otherwise specified' (ICD-9 151.9), another 'malignant neoplasm of pylorus' (ICD-9 151.1) and another 'malignant neoplasm of pyloric antrum' (ICD-9 151.2), only codes ICD-9 151.1 and ICD-9 151.2 would be retained because they are the most specific.

If at this stage only one site code remains, it is deemed to represent the primary site. If more than one remains but the only difference occurs in the fourth digit of the ICD (e.g., 151.1 and 151.2), then the site is selected from the most reliable source. For this purpose, RCC and PMH records are considered to be the most reliable source, followed by pathology records, then hospital discharge records and, finally, death certificates.

If more than one 3-digit site code remains, histology codes on each record for a patient are compared. Histology codes considered to be the same are organized into groups, according to a modification of the classification prepared by Berg (1982), as presented in Table 1. Records with a blank histology field, or in which the histology is either 'neoplasm not otherwise specified' (ICD-O M-800) or 'no microscopic confirmation of tumour' (ICD-O M-999), are included in all histology groups. In addition, records in which the histology given is 'carcinoma not otherwise specified' (ICD-O M-801) and 'carcinoma undifferentiated type not otherwise specified' (ICD-O M-802) are included with all histology groups except sarcoma, lymphoma and leukaemia. The ICD-O M codes not given in the table are considered to each have a different histology from any other, for example, 'mucoepidermoid neoplasms' (ICD-O M-843).

The OCR considers a second site of cancer in the same individual to be metastatic unless clearly shown to be otherwise. Thus the rules for reporting second primary cancers are conservative. For two different primary sites to be reported in the same individual, the sites must be different at the 3-digit ICD level and the histologies of the two sites must be in different groups, as given in Table 1. The only exception to this rule is breast cancer. Other sites rarely metastasize to the breast; if breast cancer is given as the site on a source record, then it will always be reported as a primary site, even if the histology is in the same group as that of other primary sites in the linked records. The case resolved from the other primary sites is also reported.

If different 3-digit site records have histologies in the same group according to Table 1, one or more sites are considered to be metastatic and the site reported by the most reliable source, as defined earlier, will be allocated as the primary site. If the sources are equally reliable, a broad code which encompasses all the more specific site

Table 1. Groupings of histological codes considered to be the same for allocation of site in the Ontario Cancer Registry

Alphabetical ICD-O M	Numerical ICD-O M[a]
Squamous cell carcinomas	807–808
Transitional cell carcinomas	812–813
Adenocarcinomas	814, 816, 818–823, 825–838, 857
Adnexal carcinomas	839–842
Cystic, mucinous and serous carcinomas	844–848
Ductal carcinomas	850–854
Specialized gonadal carcinomas	859, 860, 862–867
Paragangliomas and glomus carcinomas	868–871
Melanomas	872–874, 876–878
Sarcomas and other soft tissue carcinomas	880, 881, 883–886, 889–892, 899
Teratomatous carcinomas	908, 909
Blood vessel and lymphatic vessel carcinomas	912–915, 917
Osteosarcomas, chondrosarcomas and odontogenic tumours	918, 919, 922–927, 929–933
Other tumours (pinealoma, chordoma and granular cell myoblastoma)	936, 937
Gliomas	938–948
Neuroepitheliomatous tumours	949, 950
Nerve sheath tumours	954, 956
Lymphomas and Hodgkin's disease	959–966, 969–972, 975
Leukaemias	980–994

[a] Morphology codes in the *International Classification of Diseases for Oncology* (ICD-O) (WHO, 1976b)

codes from the most reliable sources is selected as representing the primary site. For example, adenocarcinomas of the transverse colon (ICD-9 153.1) and of the rectum (ICD-9 154.1), reported by equally reliable sources, would be allocated to 'large intestine, not otherwise specified' (ICD-9 153.9).

Although the OCR does not report non-melanotic skin cancer (ICD-9 173), records of this site are used, if appropriate, to override the site and histology given by less reliable sources. Thus, a pathology record with a diagnosis of cancer of skin of lip (ICD-9 173.0) will result in a non-reported case, even though the hospital record indicated a diagnosis of cancer of the lip (ICD-9 140.9) as the site of cancer.

The primary site is resolved to 'malignant neoplasm without specification' (ICD-9 199) when more specific allocation cannot be achieved. This can happen in three ways: first, if the only site recorded is 199; second, if there are two possible primaries in different organ groups with the same histology and equally reliable sources; and third, if more than one secondary site (ICD-9 196-198) is specified in the absence of a primary site.

If more than one primary site is identified, each source record is examined and associated with the most appropriate of these diagnoses, thereby aggregating all data for each site. Finally, a composite case record for each case diagnosed is created comprising the best set of diagnostic information, as determined by the computer, using all source records. Checks are made to ensure consistency of composite

Table 2. Variables retained in the composite Ontario Cancer Registry record

Record	Variable
Composite identification record	Registry identification number Names (including alternates) Sex, date and place of birth Hospital and residence codes Last known date Vital status as of last known date Cause of death from the death certificate, if deceased
Composite case record	Cancer site and histology codes Date of diagnosis Method of confirmation Residence at time of diagnosis Earliest known treatment date Hospital and RCC/PMH[a] chart numbers Hospital of diagnosis.

[a] RCC, regional cancer centre; PMH, Princess Margaret Hospital

information, (e.g., that site and histology are not in conflict, that date of death does not precede date last known alive, etc.).

The Case Resolution system is being continually improved. Cases which cannot be resolved by the rules, or which include inconsistencies, are reviewed by OCR staff and may result in subsequent modification of the computer rules. Rule changes can be encoded into the system and the entire registry file reprocessed according to the new rules.

Variables in the Ontario Cancer Registry

Three kinds of records exist in the registry: source records, composite identification records and composite case records. Source records are obtained from the external sources previously described, and represent cancer-related events. Composite identification records are created by record linkage, and represent the best identifying and demographic information on each cancer patient. Finally, composite case records are created by the Case Resolution system, and each composite case record describes an individual primary case of cancer. Most analyses use the variables in the composite records as listed in Table 2. In general, other tumour-specific variables, such as extent of disease, are not available on the source records and so are not included in the OCR.

Advantages and limitations of the Ontario Cancer Registry

The advantages of the unique system of registration in Ontario are several. The multiple sources of data combine to provide incidence data of good quality and completeness. A recent study of completeness of cancer registration in 1982, using capture-recapture methodology, estimated completeness for all sites combined as more than 95%, with a low of 91% for cutaneous malignant melanoma to a high of over 98% for deep-seated digestive organs (Robles *et al.*, 1988). When data from

several sources are present for a given diagnosis, the OCR system takes advantage of the known strengths of each particular source to select the best information for each variable.

The multiple sources generate a 'patient profile', in that data from all hospital discharges related to the cancer diagnosis are stored in the OCR, so that length of stay and other data which are valuable to health planners are available. The patient profile also readily permits identification of multiple cancers in the same patient.

This unique method of registration is relatively inexpensive, an important feature in a jurisdiction the size of Ontario.

Deaths from causes other than cancer are regularly linked to patient records along with cancer deaths. In addition, linkage with the Ontario Motor Vehicle Driver Licence file for 1964 to 1984 diagnoses has permitted positive identification of vital status for most cases who have neither died nor sought medical care for some time. These two linkages improve the quality of the OCR (e.g., for date of birth and residence), and permit generation of survival statistics by age, sex, and site.

Finally, the use of computerized linkage and Case Resolution systems ensures that records are processed in a consistent fashion. If the rules for allocating primary site are enhanced, the quality of the incidence data for the entire period of the OCR can be improved, as the complete registry data-base (more than six million records on more than 500 000 cases) can be processed by the improved system. In addition, the impact of different rules for multiple primaries on incidence rates could be assessed by processing the entire registry file through two separate case resolution programs and comparing the results.

There are, however, some limitations of the system as well. These are primarily related to reliance on, and therefore limited control over, the type, quality and flow of input data. For some data sources, coding of site, histology and residence is decentralized to hospital medical record departments (hospital separation reports) and the Office of the Registrar General (death certificates). Thus, the OCR has no control over the quality of data from these sources, both in coding and number of records received. Nevertheless, approximately 50% of cancer cases in the province are eventually referred to the RCCs or the PMH, where data quality and uniformity of coding are ensured.

Changes in any of the input data sources can affect registration, so that time trends in cancer incidence may be subject to artefacts related to changes in or problems with sources. For example, in 1978, changes in the administrative arrangements concerning provision of hospital discharge data to the MOH resulted in increased numbers of hospital discharge reports being received by the OCR, thereby producing a sudden increase in incidence rates. Also, throughout the 1970s the annual number of hospital pathology reports voluntarily submitted to the OCR increased dramatically. Thus, during these years, the OCR would be expected to include increasing numbers of patients reported solely by pathology departments and never admitted to hospital or referred to the PMH or the RCCs. For some sites, particularly those in which admission to hospital as an in-patient is not common, artefactual increases may thus be evident throughout this period. Such effects are likely to become less frequent as the OCR matures.

Another problem with dependence on outside data sources is the resultant delay in generation of incidence data. HMRI processes all hospital discharge data for a fiscal year at the end of the year; only afterwards are data passed to the OCR. Since hospital discharges comprise the major registry source, linkages cannot begin until these data are received. Nine months of processing are required at the OCR to add one year of data. Therefore incidence data for a particular year are not available until 18–24 months after the close of that year.

Further problems, occurring from the use of data generated for purposes other than cancer registration, are lack of complete demographic/geographic/tumour-specific information. For example, data are not available on clinical stage at diagnosis except for a proportion of patients seen at the RCCs or PMH. Municipality of residence is not available historically for many patients or, currently, for those for whom pathology reports only are received. Age and exact date of birth, an important linkage variable, are sometimes missing, particularly from pathology reports. This contributes to the relatively high proportion (0.06%) of cases reported with unknown age.

In addition, it is likely that there is some over-reporting in the OCR. Although a large proportion of duplicate records (where records for the same individual have not been brought together by the linkage system) are eliminated by the internal linkage process using pockets other than NYSIIS, it is estimated that 0.2% duplicates remain. This is an insoluble problem in a province where individuals do not have unique identifiers. However, the magnitude of the problem of over-reporting, owing to failure to correctly link all records, is much less than it would have been if a completely manual linkage were performed. Comparison of data using the present system applied to 1965–66 incidence data with results of the original manually linked data for these years (MacKay & Sellers, 1970, 1973), demonstrates an 11% reduction in the number of cases.

Conclusion

The OCR in its present form is a new registry, since incidence data for 1972–1976 were only produced in 1983, and for 1977–1982 in 1984. However, now that the registration techniques are well established, incidence data are added annually and are available about 18–24 months after the close of a year. Efforts are continually being made to shorten this interval, and with increasing computerization of hospital discharge and pathology reports at source, production of more timely incidence data will become feasible. The OCR is always investigating new sources of data, such as cytopathology and haematology reports, to augment the other routine data sources and thereby to improve both completeness and quality of the OCR. In addition, the linkage and Case Resolution systems are constantly being improved and streamlined, and the quality and timeliness of OCR incidence data will improve as they do. Because of the OCR's sophisticated computerized record linkage capabilities, computerized data sources outside the health care system (such as the Ontario Motor Vehicle Driver Licence file) can be linked to the OCR to improve demographic and last status variables.

The innovative method of cancer registration using computer technology has

resulted in a cancer registry of good quality, where the proportion of histologically verified cases exceeds 85%, and death certificate only registrations comprise about 2% of cases. As more computers are introduced into different aspects of the health care system, the OCR's computer-based approach may prove to be the optimum technique for cancer registration in the future.