Chapter 5

Database on cancer survival from developing countries

R. Swaminathan

Cancer Institute (WIA)
Chennai (Madras) 600036
India

R.J. Black and R. Sankaranarayanan

International Agency for Research on Cancer
Lyon,
France

Introduction

Ten population-based cancer registries from five countries in Asia and South America have contributed data on survival for this study. The registries are as follows (Fig. 1):

Qidong, China
Shanghai, China
National Cancer Registry, Cuba
Bangalore, India
Barshi, India
Bombay, India
Madras, India
Rizal, Philippines
Chiang Mai, Thailand
Khon Kaen, Thailand

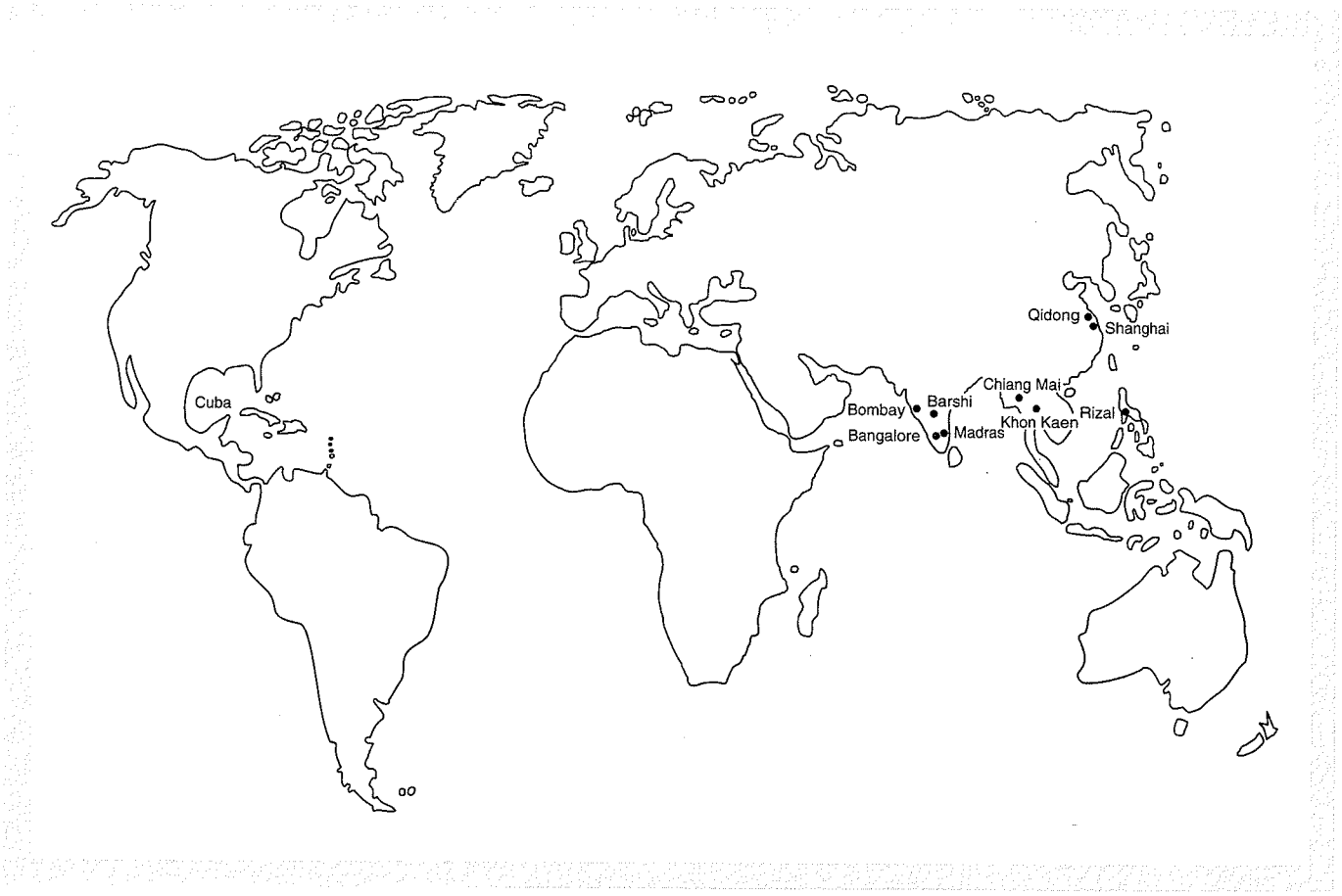

Figure 1. Map showing 10 study locations

Table 1. Cancer registration, coding and follow-up practices in participating registries

Registry, Country	Cancer registration			Coding practices		Details of follow-up	
	Period included	Method and population	Year started	Topography	Morphology	Closing date	Methods
Qidong, China	1982–91	Active and passive methods, predominantly rural	1972	ICD-9 (4 digits)	Local codes	31-Dec-94	Predominantly passive - matching with death certificates - scrutiny of medical records - house visits
Shanghai,China	1988–91	Passive notification, rural and urban	1963	ICD-9 (4 digits)	Local codes	31-Dec-94	Predominantly passive - matching with death certificates - scrutiny of medical records - house visits
Cuba	1988–89	Passive notification–voluntary till 1986, compulsory thereafter; entire country covered	1964	ICD-O I ed. (4 digits)	ICD-O I ed. (5 digits)	31-Dec-94	Predominantly passive - matching with death certificates - matching with national identity register - scrutiny of medical records - postal enquiries
Bangalore, India	1982–89	Active data collection by registry staff, totally urban	1982	ICD-9 (3 digits)	ICD-O I ed. (6 digits)	31-Dec-93	Predominantly active - matching with death certificates - repeated scrutiny of case records - postal/telephone enquiries/house visits
Barshi, India	1988–92	Active data collection by registry staff, totally rural	1987	ICD-9 (4 digits)	ICD-O I ed. (6 digits)	31-Dec-95	Predominantly active - matching with death certificates - repeated scrutiny of case records - postal/telephone enquiries/house visits
Bombay, India	1982–86	Active data collection by registry staff, totally urban	1963	ICD-O I ed. ICD-9 (4 digits)	ICD-O I ed. (6 digits)	31-Dec-93	Predominantly active - matching with death certificates - repeated scrutiny of case records - postal/telephone enquiries/house visits
Madras, India	1984–89	Active data collection by registry staff, totally urban	1982	ICD-O I ed. ICD-9 (4 digits)	ICD-O I ed. (6 digits)	31-Dec-93	Predominantly active - matching with death certificates both cancer and not cancer as cause - repeated scrutiny of case records - perusal of area health registers - postal/telephone enquiries/house visits
Rizal, Philippines	1987	Passive notification till 1978 - active since then, mainly urban	1974	ICD-O I ed. ICD-9 (4 digits)	ICD-O I ed. (5 digits)	31-Dec-93	Predominantly passive - matching with death certificates - matching with case finding lists - scrutiny of medical/health records - enquiries with attending physician
Chiang Mai, Thailand	1983–92	Active data collection by registry staff, rural and urban	1963 Hospital-based till 1986. Population-based since then - retrospective data collected from 1983	ICD-O I ed. (4 digits)	ICD-O I ed. (4 digits)	30-Jun-94	Predominantly active - matching with death certificates - repeated scrutiny of medical records - postal enquiries/house visits
Khon Kaen, Thailand	1985–92	Active and passive methods, rural and urban	1984 Hospital-based till 1988. Population-based since then - retrospective data collected from 1984	ICD-O I ed. (4 digits)	ICD-O I ed. (5 digits)	31-Dec-95	Predominantly active - matching with death certificates - scrutiny of hospital records - postal enquiries/house visits

Table 2. Cancers included and variables provided for analysis

Registry, Country	Cancer sites included : ICD-9	Variables provided by the registry
Qidong, China	147, 150–5, 157, 162, 174, 180, 188, 191–2*, 203–8	Identification number, age at incidence date, sex, date of birth, most valid basis of diagnosis, incidence date, primary site of cancer, dates of death/last follow-up, vital status, survival time
Shanghai, China	140–208	Identification number, age at incidence date, sex, most valid basis of diagnosis, incidence date, primary site of cancer, dates of death/last follow-up, vital status, survival time
Cuba	140–1, 143–6, 153–4, 162, 174,180, 182–3, 185, 200–8	Identification number, age at incidence date, sex, date of birth, most valid basis of diagnosis, incidence date, primary site of cancer, clinical extent of disease, morphology, dates of death/last follow-up, vital status, survival time
Bangalore, India	174, 180, 200–8	Identification number, age at incidence date, sex, most valid basis of diagnosis, incidence date, primary site of cancer, tumour stage/clinical extent of disease, histology, dates of death/last follow-up, vital status
Barshi, India	180	Identification number, age, sex, religion, most valid basis of diagnosis, incidence date, primary site of cancer, tumour stage,morphology,treatment, dates of death/last follow-up, vital status, survival time
Bombay, India	174, 180	Identification number, age at incidence date, sex, marital status, mother tongue, religion, literacy, most valid basis of diagnosis, incidence date, primary site of cancer, morphology, dates of death/last follow-up, vital status
Madras, India	140–1, 143–6,148,150–1, 157, 161–2, 174, 180, 188, 200–2, 204–8	Identification number, age at incidence date, sex, marital status, mother tongue, religion, literacy, most valid basis of diagnosis, incidence date, primary site of cancer, morphology, dates of death/last follow-up, vital status, survival time
Rizal, Philippines	143–5, 151, 153–5, 162, 174, 180, 185, 204–8	Identification number, age at incidence date, sex, date of birth, most valid basis of diagnosis, incidence date, primary site of cancer, clinical extent of disease, morphology, dates of death/last follow-up, vital status
Chiang Mai, Thailand	140–208	Identification number, age at incidence date, sex, most valid basis of diagnosis, incidence date, primary site of cancer, clinical extent of disease, morphology, treatment, dates of death/last follow-up, vital status
Khon Kaen, Thailand	140–208	Identification number, age at incidence date, sex, marital status, ethnicity, religion, most valid basis of diagnosis, incidence date, primary site of cancer, clinical extent of disease, morphology, dates of death/last follow-up, vital status, survival time

* Includes benign and unspecified neoplasms, number not known

Data on the following variables were requested from each participating registry:

- identification number
- sex
- date of birth
- socioeconomic factors (marital status, mother tongue, religion, ethnicity, education, socioeconomic status, etc.)
- incidence date
- age at incidence date
- most valid basis of diagnosis of cancer
- clinical extent of disease before treatment/tumour stage
- primary site of cancer (ICD-O, ICD-9)
- morphology
- date of death/last follow-up
- vital status at this date (alive/dead/lost to follow-up).

Tables 1 and 2 give the details of the study period, cancers studied, variables provided, cancer registration and follow-up methods, and coding practices followed for each of the registries. Since all these registries were population-based, they were asked to send data on *all* incident cases for the period under study, not merely for the subset of cases for which follow-up information was available. This was done mainly to evaluate the usual indicators of data quality (proportion of cases with a histological verification of cancer diagnosis, proportion of cases registered on the basis of death certificate only) and to permit the entire dataset to be subjected to standard validation checks.

Study period

The participating registries had been in operation for varying periods of time. The years for which follow-

up information was available were even more varied. For this reason, it was decided to use the maximum period of data available from each registry, rather than imposing a single time period on all. The periods of registration under study were all between 1 January 1982 and 31 December 1992.

Study material

Not all the participating registries could provide follow-up information on all the cancer sites registered during the study period. Chiang Mai and Khon Kaen from Thailand and Shanghai from China had follow-up information for all cancer sites. Data were available only for selected cancer sites in the other registries: for the most part, these were the most common cancers in their respective regions. Follow-up of breast and cervical cancers in females and lung cancer and other tobacco-related cancers in both sexes had been carried out in most registries (Table 2).

All the incident cancer cases (in the sites chosen for analysis) were included in the study. Only invasive cancers were included. No distinction was made between first and subsequent primary cancers in the same individual. However, in most developing countries, second and subsequent primaries constituted a negligible proportion of the total.

Primary site of cancer

Data on the tumour site had been coded in accordance with the *International Classification of Diseases for Oncology,* First Edition (ICD-O) (WHO, 1976) by all the registries, and some had simultaneously coded the site in accordance with the *International Classification of Diseases, Ninth Revision* (ICD-9) (WHO, 1978). Conversion to ICD-9 codes was done wherever necessary. Details of the diagnostic categories used are shown in Table 3. Only categories with at least 25 cases were considered for analysis.

Morphology

Morphology (and behaviour), coded in accordance with the *International Classification of Diseases for Oncology,* First Edition (WHO, 1976), were available for all the datasets except the two from China, which had used local codes (Chinese characters) for morphology.

Tumour stage/clinical extent of disease

Data on tumour stage, classified by TNM (tumour-node-metastasis) stage categories, were not routinely available in the population-based cancer registries participating in this study. However, data on the clinical extent of disease before treatment were available in all the registries (at least for selected sites) except the two from China. The criteria followed by the participating registries are shown in Table 4. There is bound to be variation in the accuracy of this information between registries, as it depends on the extent of investigative procedures and on registration practices. Because of this variability, comparison of survival estimates by clinical extent of disease was confined to selected sites within individual registries.

Index date

There are several possible starting dates for calculating survival. The one most widely available in population-based cancer registries is the incidence date. This date, as provided by the registries, was taken as the index date for this study. A review of the definitions used by the registries for coding the incidence date did not reveal any substantial variation. Such a variation might lead to minimal differences in short-term survival (<2 years) and will be less evident in long-term survival (Berrino *et al.,* 1995). The index dates in this study ranged from 1 January 1982 and 31 December 1992.

Closing date

The closing date, or date of last follow-up, varied between registries and ranged between 31 December 1993 and 31 December 1995. Each patient's vital status was classified as dead, alive or lost to follow-up as on the closing date.

Survival time

This was calculated as the time (in months) between the index date and the date of death from any cause *or* date of loss to follow-up *or* the closing date, whichever was earliest. The date of loss to follow-up was assumed to be the middle of the year/month if only the year/month was known; it was taken to be 31 December of the calendar year if it occurred in the same calendar year as diagnosis and the precise date of loss to follow-up was not known.

Data quality indicators

Two indices of data quality were calculated: (1) the percentage of cases with histological verification of cancer diagnosis and (2) the percentage of cases registered on a death certificate only (DCO) basis. Other aspects of data quality, especially those

concerned with the completeness of ascertainment and follow-up, are discussed in the chapters dealing with individual registries.

Exclusions

Two categories of cases were excluded from the analysis.

(1) Cases based on DCO registrations (i.e. ones for which no information prior to death certificate could be traced) and cases first identified at autopsy. The percentage of cases so excluded ranged from 0% to 42.7%, as specified in the registry results chapters.

(2) Cases for which no follow-up information was available after the incidence date. The percentage excluded for this reason ranged from 0% to 11.9% in different registries.

Study database

The following variables were included in the database created for analysis.

1. Registry identification number (two-digit code for each registry).
2. Sex (1: male; 2: female).
3. Age at incidence date.
4. Primary site of cancer (ICD-9: 140-208).
5. Morphology (ICD-O, first edition, where available).
6. Clinical extent of disease (1: localized; 2: regional; 3: distant metastasis; 4: unknown; where available).
7. Incidence date (mm/dd/yy).
8. Date of death/closing date/date of loss to follow-up (mm/dd/yy).
9. Vital status of patients at this date (0: dead, 1: alive, 2: lost to follow-up).
10. Survival time (in months).

Validation checks

A set of validation checks was performed prior to survival analysis. A list of the checks undertaken, with the range of errors encountered among the registries, is given in Table 5. The CHECK program (Parkin *et al.*, 1994) was used to detect inconsistencies in age, site and histology combinations. The CONVERT program (Ferlay, 1994) was used to convert primary site codes in ICD-O, first edition (WHO, 1976) and ICD-O, second edition (WHO, 1990) to ICD-9 (WHO, 1978) wherever necessary. Age at diagnosis was recalculated whenever the date of birth was available.

A list of any potential errors was returned to the registry for clarification and correction. The validation checks were then repeated on the revised data. Tables showing the proportion of cases finally excluded from the study by site are given separately in the chapters dealing with the individual registries.

Overview of study material and methods

All new incident cancers (all sites/selected sites) registered in each registry

Cases registered based on DCO/autopsy Cases without follow-up information Cases rejected on validation	=	Excluded from analysis
Cases accepted on validation - with complete follow-up (alive/ dead status known at closing date) - with partial follow-up (censored prior to closing date)	=	Included in the analysis
Case registration period	=	varies between registries (1 Jan 1982–31 Dec 1992)
Index date for the analysis	=	Date of incidence
Closing date of the analysis	=	varies between registries (31 Dec 1993–31 Dec 1995)
End-point of the analysis	=	Death of the patient irrespective of cause
Observed and relative survival rates	=	Computed by Hakulinen's method (Hakulinen, 1982, Hakulinen *et al.*, 1994)

Table 3. Description of sites chosen for survival analysis

ICD-9 code	Title	ICD-9 description
140	Lip	Lip
141	Tongue	Tongue
142	Salivary gland	Major salivary glands
143–5	Oral cavity	Gum, floor of mouth, unspecified parts of mouth
146	Oropharynx	Oropharynx
147	Nasopharynx	Nasopharynx
148	Hypopharynx	Hypopharynx
150	Oesophagus	Oesophagus
151	Stomach	Stomach
152	Small intestine	Small intestine
153	Colon	Colon
154	Rectum	Rectum, rectosigmoid junction, anal canal and anus
153–4	Colorectal	Colon and rectum
155	Liver	Liver
156	Gallbladder	Gallbladder
157	Pancreas	Pancreas
161	Larynx	Larynx
162	Lung	Bronchus, trachea and lung
170	Bone	Bone
171	Connective tissue	Soft and connective tissues of all regions
172	Skin melanoma	Melanoma of the skin of any part
173	Skin non-melanoma	Non-melanomatous skin of any part
174	Breast	Female breast
180	Cervix	Cervix uteri
182	Corpus uteri	Corpus uteri
183	Ovary	Ovary and other uterine adnexa
184	Vagina	Vagina, vulva and unspecified female genital organs
185	Prostate	Prostate
186	Testis	Testis
187	Penis	Penis, scrotum and unspecified male genital organ
188	Bladder	Urinary bladder
189	Kidney	Kidney, urethra and other urinary organs
191–2	Brain, nervous system	Brain and other central nervous system
193	Thyroid	Thyroid
201	Hodgkin's disease	Hodgkin's disease
200,202	Non-Hodgkin lymphoma	Lymphosarcoma and non-Hodgkin lymphoma
203	Multiple myeloma	Multiple myeloma
204	Lymphatic leukaemia	Acute, chronic and other lymphatic leukaemia
205	Myeloid leukaemia	Acute, chronic and other myeloid leukaemia
204–8	All leukaemia	All types of leukaemia
195–9	Primary site uncertain	Ill-defined sites, primary site unknown

Table 4. Criteria used for classification of clinical extent of disease

Category	Description
Localized	Tumour confined to the organ of origin, without invasion into the surrounding tissue/organ and without involvement of any regional or distant lymph nodes or organs
Regional	Tumour not confined to the organ of origin, with invasion into the surrounding tissue/organ, with or without the involvement of the regional lymph nodes and not involving the nonregional lymph nodes or organs
Distant metastasis	Tumour involving the nonregional lymph nodes or distant organs
Unknown	The above information is unknown

Table 5. Validation checks and range of errors detected among registries

Validation check	Range (%) of errors
Age or sex unknown	0.0–1.40
Date of diagnosis — out of range	0.0–0.02
Date of death/last follow-up — out of range	0.0–0.02
Cases with negative duration of survival time	0.0–0.10
Primary site code — out of range	0.0–0.03
Unlikely age and site combination	0.0–0.04
Unlikely sex and site combination	0.0–0.10
Conversion error of site code from ICD-O to ICD-9	0.0–0.01
Histology codes — out of range	0.0–0.60
Unlikely site and histology combination	0.0–0.70
In situ cancers	0.0–0.30
Invalid vital-status codes	0.0–0.30

References

Berrino, F., Sant, M., Verdecchia, A., Capocaccia, R., Hakulinen, T. & Estève, J., eds. (1995) *Survival of Cancer Patients in Europe: the EUROCARE Study* (IARC Scientific Publications No. 132). Lyon, International Agency for Research on Cancer

Ferlay, J. (1994) *ICD Conversion Programs for Cancer* (IARC Technical Report No. 21). Lyon, International Agency for Research on Cancer

Hakulinen, T. (1982) Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, **38**, 933–942

Hakulinen, T., Gibberd, R., Abeywickrama, K.H. & Soderman, B. (1994) *A Computer Program Package for Cancer Survival Studies, Version 2.0.* Tampere, Finnish Cancer Registry/University of Newcastle, Australia

Parkin, D.M., Chen, V.W., Ferlay, J., Galceran, J., Storm, H.H. & Whelan, S.L. (1994) *Comparability and Quality Control in Cancer Registration* (IARC Technical Report No. 19). Lyon, International Agency for Research on Cancer, pp. 61–65

WHO (1976) *International Classification of Diseases for Oncology,* First Edition. Geneva, World Health Organization

WHO (1978) *International Classification of Diseases, Ninth Revision.* Geneva, World Health Organization

WHO (1990) *International Classification of Diseases for Oncology,* Second Edition. Geneva, World Health Organization