# Toward Optimizing Data Structures in Cancer Registries

Anokhi J Kapasi, Varinder P Singh, Sharmila A Kamani, **Judith K Jones**
DGI, LLC, Fairfax, VA, USA E-mail: info@bridgetodata.org

## BACKGROUND

- Cancer registries are critical tools to systematically monitor incidence, geographic distribution, and characteristics of many types of cancers.
- There is no consensus for data collected in cancer registries.
- A standard group of data elements and format would allow valuable comparisons among registries worldwide.
- B.R.I.D.G.E. TO DATA® (www.bridgetodata.org), a resource of >250 database profiles worldwide (currently 35 countries), can contribute to optimization of these data collected for registries.

## OBJECTIVES

To describe and characterize cancer registries by ***frequency*** and ***types*** of data to identify gaps and opportunities for enhancing current registries and building future ones.

## METHODS

**Box 1.** Using the following search criteria, a search was conducted in B.R.I.D.G.E. TO DATA® to identify registries that collect cancer data:
***Database Type* = *Registry*; and *Cancer Data* = *Yes*** (**Figure 1**).

**Box 2.** One hundred ninety-nine (199) profiles matched at least 1 criterion (**Figure 2**): **(a)** 46 registry profiles matched both criteria; **(b)** Search results were further narrowed by excluding 24 profiles of non-cancer registries.

**Box 3.** Sixty-nine (69) of the 75 relevant data fields used in the B.R.I.D.G.E. structured profiles were compared among the 22 cancer registries (**Table 1**).

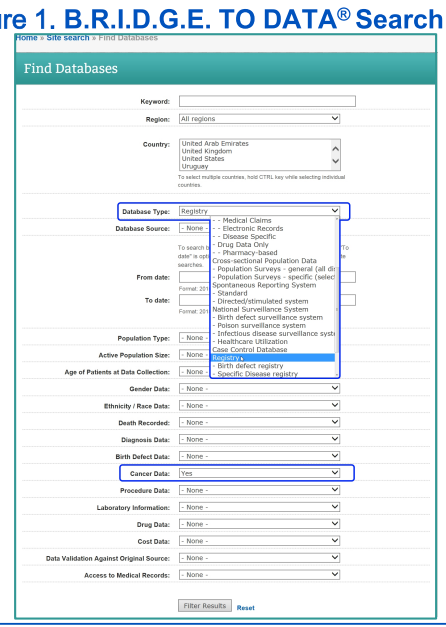
**Box 4.** For each profile, frequency counts of data field usage in the registry were obtained (e.g., *Date of Birth* captured or not).

**Box 5.** Data fields were grouped based on the frequency of usage among the 22 cancer registries.

**Box 6.** The frequency of usage categories were:
- ***Core Data Fields*** with similar frequency of use among all 22 registries;
- ***Additional Data Fields*** present in >50% of registries;
- ***Infrequently Used Data Fields*** present in ≤50% of registries.

**Box 7.** Data fields were **further subcategorized** based on similarity in the type of data captured (e.g., *Diagnoses* captured by the same coding system).

### Figure 1. B.R.I.D.G.E. TO DATA® Search Page




### Table 1. Examples of Data Fields Used in Profiles (by Category)

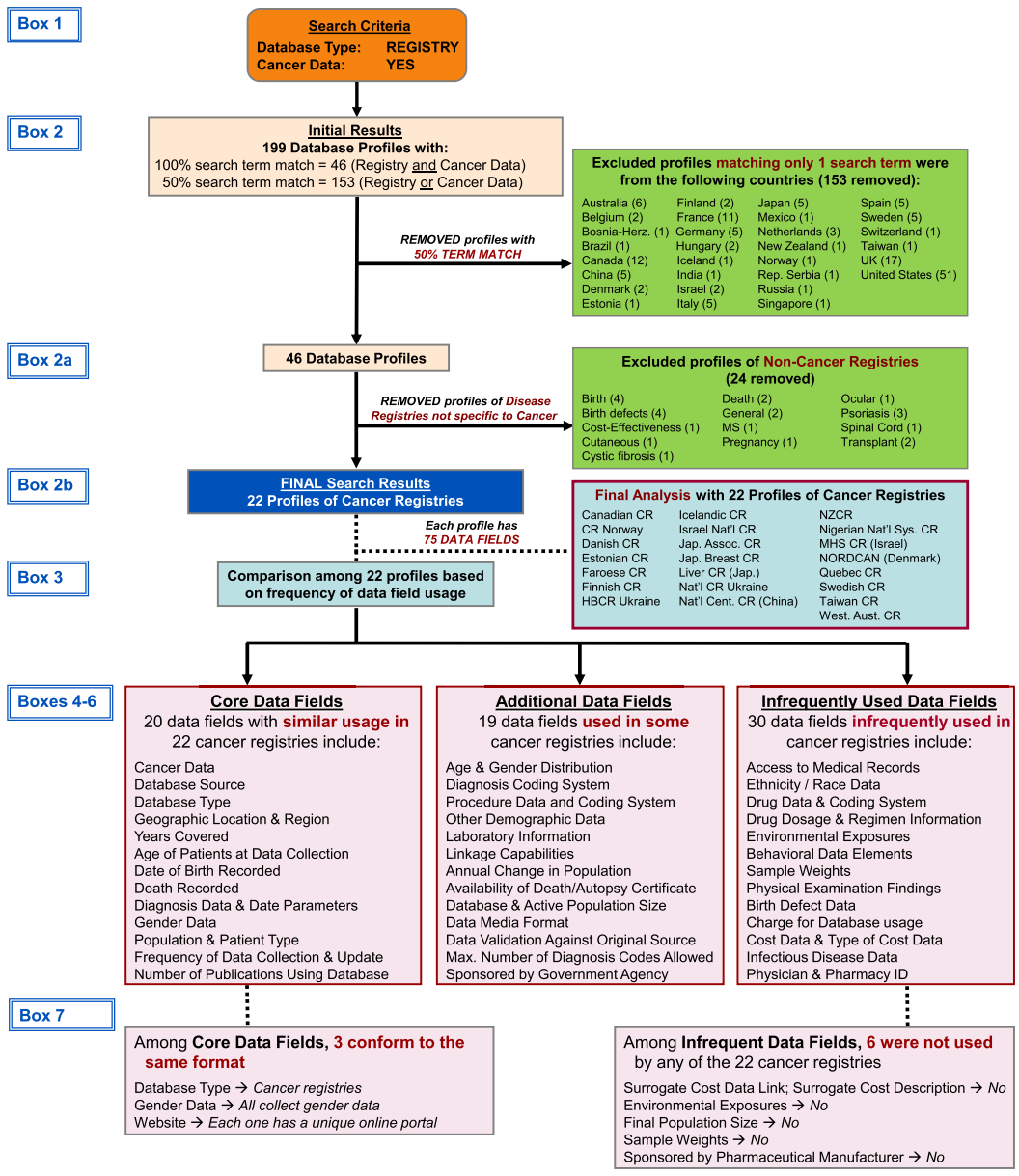
| Category | Data Fields |
|---|---|
| Summary | Database description, Database source, Years covered, Population type, Date of last update |
| Population Dynamics | Population size, Sample weights – Extrapolation factors |
| Demographic Data | Age, Gender, Date of birth, Death recorded, Other demographic data |
| Physician & Practitioner Info | Physician ID & Specialty, Pharmacy ID |
| Diagnoses/Signs & Symptoms | Diagnosis data, Diagnoses coded (coding systems), Max. number of codes, Physical exam findings, Environmental exposures, Behavioral data elements |
| Procedures | Procedure data, Procedures coded (coding systems), Laboratory information |
| Drug Information | Drug data, Drug dosage, Drug coding system(s), Additional drug information |
| Economic Data | Type of cost data (if applicable) |
| Validation & Linkage | Data validation, Access to medical records, Linkage to other databases |
| Administrative Data | Database contact data, Database usage restrictions, References of studies using/describing the database |

## RESULTS

- These 22 cancer registries included nationally representative populations, were mostly initiated before year 2000, and were systematically updated from multiple data sources (**Figure 2**).
- Cancer diagnoses were predominantly recorded with ICD-10/ICD-O-3 codes and variably included diagnosis date & histology/ pathology/ staging data.
- Twenty (20) core data fields were used by all cancer registries: 3 fields captured identical data (e.g., gender) while 17 had variable data (e.g. patient type, date of birth format, death data, source).
- Another 19 fields were utilized by the majority of registries (e.g., procedures, lab data).
- About half captured cause of death (11/22) and date of death (12/22), primarily obtained from autopsies/death certificates.
- Of 30 infrequent fields (e.g., sociodemographic, physical exams, drugs), 6 were not utilized by any registry.
- Although drug data (10/22) were infrequently captured, procedure data (17/22) were common.

## RESULTS – Part 2

### Figure 2. Criteria-based Search for Cancer Registries Conducted in B.R.I.D.G.E. TO DATA® (252 Database Profiles worldwide as of August 15, 2015)




### Table 3. Excerpt from B.R.I.D.G.E. TO DATA® Comparing Data Elements in 3 Selected Cancer Registries

| FIELD NAMES | Israel National Cancer Registry (INCR) (Israel) | Nigerian National System of Cancer Registries (Nigeria) | Swedish Cancer Registry (Sweden) |
|---|---|---|---|
| Database Type | Registry - Specific Disease registry (Passive cancer registry) | Registry - Specific Disease registry (NSCR is a network of population-based and hospital-based cancer registries in Nigeria) | Registry - Specific Disease registry (Cancer Registry) |
| Database Source | Passive reporting system. Reports are received from: Pathology laboratories, Hospitals (discharge summaries), Oncology clinics (new patient lists), Death certificates, Israel National Population Register (demographic information, address, vital status), Israel Central Bureau of Statistics (causes of death) | Case Report Forms | Medical Records<br>Death Certificates<br>A "clinical" report has to be sent for every cancer case diagnosed at clinical, morphological, other laboratory examinations, as well as cases diagnosed at autopsy |
| Years Covered | 1960 - Present | 2009 - Present | 1958 - 2012 |
| Population & Patient Type | General population (Patients with cancer); inpatient and outpatient | General Population (People diagnosed with cancer within the catchment area of the population-based cancer registries and cancer patients at hospitals that have hospital-based cancer registries); inpatient and outpatient | General Population (Those who are diagnosed with cancer); both inpatient and outpatient |
| Database Population Size (Range) | 0.5 - 1 Million (This refers to the number of cancer diagnoses) | <200,000 (NSCR database has information on 22,322 cancer patients) | 1 - 5 Million (2.2 Million) |
| Approximate Percentage of Participants <18 years and those >65 years | ≤18 years = 2%<br>>65 years = 50%<br>(As of 2011) | <18 years = 5% (1,119 cases, i.e., 5% of total)<br>>65 years = 23% ( 5,125 cases, i.e., 23% of total) | <18 years=0.5%<br>>65 years=76% |
| Ethnicity / Race Data | Yes (Jewish, Arab, other) | Yes, ethnic group is captured | No |
| Death Recorded | Yes (Date of death) | Yes, follow-up data on cases are recorded, and may include status of contact (Alive/Dead), cause of death and date of death. Some cancer cases are obtained from death certificates. | Yes, date and cause of death are both recorded. However, the Swedish Cancer Register does not accept notifications from death certificates. |
| Other Demographic Data | No | Yes, some additional sociodemographic information includes patient's name, institution/ward, and laboratory that data may have been obtained from | Yes, information includes: Place of residence, Personal Identification Number, domicile (county, municipality, parish), unique tumor specimen number including year when specimen was taken, site of tumor, date of migration |
| Diagnosis Data | Yes, data are available on cancer diagnoses as well as recurrence or metastatic disease (in cases in which the original stage was not metastatic). Instances of recurrence are recorded in the original case record, including date of first recurrence. | Yes | Yes, basis and date of diagnosis are collected, as well as reporting hospital and department, reporting pathology/cytology department, identification number for the tissue specimen. |
| Diagnoses Coded | ICD-O-3 | ICD-O-3; Other [Primary site of the tumor (CXX.X), morphology (5-digit structure MXXXX/X), and tumor behavior are coded using ICD-O-3.<br>Tumor stage is coded using AJCC codes (1, 2A, 2B, 3A, 3B, and 4) and the TNM classification system for specifying tumor, node, and metastasis characteristics.] | ICD-7; ICD-9; ICD-O-2; ICD-O-3 [1987 - 1992 = ICD-9; 1993 - 2004 = ICD-O-2; 2005 onwards = ICD-O-3. For the whole period (1958 - Present), codes are available as ICD-7 codes.<br>Gynecological tumors are coded according to International Federation of Gynecology and Obstetrics (FIGO), and the rest according to TNM - 6th edition. |
| Cancer Data | Yes (ICD-O-3 topography and morphology) | Yes, information is recorded on: Date of incidence (DDMMYYY), Source of diagnosis (Death certificate only, Clinical only, Clinical investigations, Specific tumor markers, Cytology/Hematology, Histology of metastasis or primary tumor, or Unknown), Primary tumor site, Morphology, Stage and TNM. | Yes, data are available on: Site & stage of tumor, basis & date of diagnosis. Stage is being collected only since 2004, though it is not recorded for brain, cranial nerves, lymphoma and leukemia. Breast cancer is the most common cancer in women and constitutes 29% of diagnosed cases. Most common cancer among men is prostate cancer, accounting for 33% of cases in 2008. Malignant melanoma and other skin cancer (except basal cell carcinoma) together constitute 14% of cancers. Colon cancer is the second most common cancer among women, third most common in men. Altogether, 8% of cases were reported to the Cancer Registry for 2008. |
| Environmental Exposures | No | No | No |
| Procedure Data | Yes, information is available on cancer-related procedures | Yes, treatment for cancer is recorded as: Surgery, Radiotherapy, Chemotherapy / Hormone therapy, and/or Other treatment. | No |
| Laboratory Information | Yes, pathology results are recorded in cases in which the pathology report is basis of diagnosis | No, type of laboratory test and results details are not captured; however, laboratory location/ID is captured. Laboratory data may be used (when available) as a source for filling out the case report form. | Yes, histological and pathology results are available |
| Drug Data | No | No | No |
| Cost Data | No | No | No |

**LIMITATIONS:** This analysis was a limited cancer registry sampling using DBs currently profiled within B.R.I.D.G.E. TO DATA®. More profiles of data sources are continually being added to B.R.I.D.G.E. Future analyses may provide a better comparison.

## CONCLUSIONS

❖ B.R.I.D.G.E. was successfully used to analyze and categorize data fields in these cancer registries and may serve as a template when improving and designing disease registries.

❖ Cancer registries commonly used ICD-10/ICD-O-3 codes for diagnoses, and included histology, pathology, and/or staging data. Half of them captured detailed information on death.

❖ Infrequently captured data may prove important for understanding diseases relating to exposures (e.g., environmental, drugs), prognosis (e.g., tumor markers), and may enhance quality of cancer studies (e.g., cost of illness studies).

**In conclusion, there is a need for consensus among cancer epidemiologists to define international categories and characteristics of data needed in all population oncology & medical product epidemiology databases.** This not only applies to the disorder but also ancillary data such as prior exposure (i.e., environmental), biomarkers, therapies, procedures, and geographic/ethnic distribution.

**This study was self-funded by DGI, LLC**

**31st International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Boston, USA. August 26, 2015 [#978]**

# Toward Optimizing Data Structures in Cancer Registries

Anokhi J Kapasi, Varinder P Singh, Sharmila A Kamani, **Judith K Jones**

DGI, LLC, Fairfax, VA, USA

Presented at the 31st International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Boston, Massachusetts, USA. August 26, 2015 [#978]


DGI Center for Health Research and Education, LLC
9302 Lee Highway, Suite 700 • Fairfax, VA 22031 USA
www.bridgetodata.org T: +1 571 490 8400 info@bridgetodata.org