

## RESEARCH ARTICLE

# Coverage, Density and Completeness of Sources used in Tehran Metropolitan Area Cancer Registry: According to the Data of Esophageal Cancer, 2003-2007

Abbas Aghaei<sup>1\*</sup>, Farid Najafi<sup>2</sup>, Alireza Mosavi-Jarrahi<sup>3</sup>, Toraj Ahmadi-Jouibari<sup>4</sup>

### Abstract

**Background:** The completeness of cancer registration is a major validity index of any reported cancer incidence. The present study aimed to evaluate the esophageal cancer incidence registered in the Tehran Metropolitan Area Cancer Registry. **Materials and methods:** The data on esophageal cancer abstracted from three sources of 1) pathology departments, 2) medical records, and 3) death certificates during 2003 till 2007 were utilized. The completeness of the data sources were evaluated using coverage (defined as the proportion of a community population with esophageal cancer identified by the source) and density (defined as the proportion of non-empty fields of the data by source). **Results:** A total 1,404 cases of esophageal cancer were reported for the duration of the study. Pathology provided 771, medical records 432, and death certificates 609. The coverage was 0.55 for pathology, 0.31 for medical records, and 0.43 for death certificates. The respective density values were 0.82, 0.96 and 0.98, respectively. Pathology (0.45) was the most complete source followed by medical records (0.42), and death certificates (0.29). **Discussion:** A low degree of completeness dictates putting more effort into case finding plus abstracting data more thoroughly.

**Keywords:** Cancer-registry - source coverage - distribution density - completeness - esophageal cancer - Iran

*Asian Pacific J Cancer Prev*, 13, 3617-3619

### Introduction

Recent development in health related technology has the number of data registries for different health condition such as cancers, cardiovascular diseases and other non-communicable and communicable diseases. Most of such registries are based on available data in health system. Such data are mostly those which are not initially collected and managed for the especial goals of an ongoing research (Hearst and Hulley, 1988). They are often collected for a systematic management of an health condition or evaluating all health activities related to that condition which might need to collect different set of data (Sorensen, 1982; Sorensen et al., 1983). In fact using the available data result in saving time and cost. In addition the large sample size used for collection of such data increase the generalizability of result and reduces the probability of information bias by not asking directly from the patient. However, using such data when the process of collecting data is not under direct supervision of an researcher may decrease the validity and accuracy. Among the limitations, it can be mentioned that controlling the collection method and quality is not under the supervision of the investigator and evaluating its validity is sometimes impossible (Hearst

and hulley, 1988; Roose et al., 1990). Completeness and accuracy of available data are the most considered when using these data (Goldberg et al., 1980; Stone, 1986). Data quality and validity of registry systems which use 2 or more data sources is directly related to quality and validity of the used sources.

Population-based cancer-registries are valuable sources for public healthcare and researches because of providing important information regarding cancer. The extend of reliance on these data is related to quality and validity of data (Elbasmi et al., 2010); however, they rarely report such information (Bullard et al., 2000) since more accessibility and valuation about the collected data is needed. In addition further investigation about the validity and completeness of such data need to be followed by looking other sources. In case of cancer registry some activities such as reabstracting, measuring the values of death certificate notification (DCN) and death certificate only (DCO), comparing the registered data with available data in other data banks will be needed (Chen et al., 1994).

In Iran the first activities in order to organize cancer reporting was started in 1956 when Cancer Society in Tehran University was founded (Habibi 1985). In 1970, observing a high incidence of esophageal cancer in

<sup>1</sup>Department of Epidemiology, <sup>2</sup>Vice chancellor of Research and Technology, <sup>4</sup>Department of Internal Medicine, Kermanshah University of Medical Science, Kermanshah, <sup>3</sup>Department of Epidemiology, The Cancer Research Center of the Cancer Institute, School of Public Health, Shaheed Beheshti University of Medical Sciences, Tehran, Iran \*For correspondence: Aqaei.a@gmail.com, aqaei.a@kums.ac.ir

**Table 1. Distribution Density for Sources used in TMACR for Esophageal Cancer, 2003-2007**

Source	All cases	Name	Surname	Father's name	Date of diagnosis	Sex	Address	Birth date
Death certificates	609	607 (0.997)	575 (0.994)	597 (0.980)	609 (1)	609 (1)	527 (0.865)	605(0.993)
Pathology reports	771	770 (0.999)	771 (1)	366 (0.475)	771 (1)	771 (1)	292 (0.379)	706(0.916)
Medical record of hospitals	432	432 (1)	432 (1)	373 (0.863)	432 (1)	432 (1)	381 (0.882)	410(0.949)

sidelines of Caspian Sea persuaded the investigators to sustain the first population-based cancer-registry center (Kmet, 1972). By now, population-based cancer-registry is only implemented in Golestan, Ardabil, Tehran, Semnan and Kerman provinces in which none has provided a report of registry quality control (Mohagheghi and Mosavi-Jarrahi, 2010; Zendedel et al., 2010).

This study aimed to evaluate the sources used in Tehran Metropolitan Area Cancer Registry (TAMCR) regarding the coverage, data distribution density and completeness in addition to provide a new method for quality control of data registration in population-based cancer registries.

**Materials and Methods**

TMACR uses three main sources of pathology reports, medical records of hospitals and death certificates for data registration (mohagheghi et al., 2009). In the present study, we evaluated all of the esophageal cancer data which have been actively or passively reported by the three mentioned sources during 2003-2007. In order to track deaths occurred in 2008 for incident cases of previous years, the death data in this year were also evaluated. All ata evaluated in regard to coverage, density and completeness which was introduced by Felix Naumann and Johann Christoph Freytag in (Naumann et al., 2000: 2004).

*Coverage*

Coverage of a source (e.g. pathology reports) is defined as the proportion of a community population (here all patients with esophageal cancer) who have been identified by the source,  $C(s)=S/P$ .

In which S is the number of the reported cases by the considered source, P is the community population (all of the patients with esophageal cancer reported to cancer registry) and C(S) is the coverage of the considered source. If it is assumed that all of the available cases are identified and each is at least registered by one source, then p is resulted from the total registered cases (regarding the common cases) by the sources,  $p=s_1 U s_2 U ... U s_n$ .

*Density*

Density is the proportion of complete (non-empty) fields of the provided data by a source. Generally, there might be many empty fields among the provided data of a source; for example, among the main fields needed for registering a case of esophageal cancer, address or date of birth might be missing. For density, two parts are respectively calculated: density of distribution and source.

a) Distribution density: this part includes distribution for each of the considered fields in a source:  $d_s(t_a)=\{t \in S | t_a \neq \perp\}/S$ . While t includes the required fields for registration (e.g. date of diagnosis of cancer),  $t_a$  is the number of complete

cases in a field and  $\perp$  is the unregistered cases in the source S. In fact for each cancer registry, the importance of each field is different. For instant, in cancer registry, date of diagnosis is one of the most important fields for determining incidence and prevalence in different time intervals. In TMACR, the fields of name, surname, father's name, date of diagnosis, sex, address and birth date are classified as the main fields.

b) Source density: density of a source is the mean of all distribution densities for each source,  $d(s)=1/A \sum_{t_a \in A} d_s(t_a)$ . In which A is the number of the considered fields in the source S.

*Completeness*

Using the criteria of coverage and density, another criterion as completeness can be defined foe each sources in a registry system,  $C(S)=c(S).d(s)$ .

**Results**

During 2003-2007, 1,404 new cases of esophageal cancer were reported to TMACR. The coverage of death certificates, pathology reports and medical records of hospitals was 0.434, 0.549 and 0.308, respectively.

In Table 1, the distribution density of the fields has been presented for each of the sources, separately; the density of death certificates, pathology reports and medical records of hospitals was 0.976, 0.824 and 0.956, respectively.

According to the values for the coverage and density of the sources, the completeness of death certificates, pathology reports and medical records of hospitals was 0.424, 0.452 and 0.294, respectively.

**Discussion**

Regarding the sources coverage, death certificates, pathology reports and medical records of hospitals had ordered from first to third rank, respectively. For the purpose of this study, we did not investigate the sensitivity and specificity of each sources and therefore the precision and accuracy of each of the sources need to be discussed in details in future studies.

In the distribution density, while the fields of diagnosis date, sex, name and surname had a density of 1 or approximately 1, the field of address in each of the three sources had the lowest density compared with other fields. It is worth mentioning that; because the diagnostic and therapeutic centers under the coverage of TMACR are referral, the low density of the address field among the reports of the used sources makes it difficult to claim that the reported cases to the center belongs to the covered population.

According to the goals of each of the registry systems, a specific weight can be considered for different fields

to have a better comparison between the sources. For example, the registry systems aiming to compare the values between males and females, different jobs, various races, birth subgroups and etc. consider a different importance scale for each of these fields (Elbasmi et al., 2010; Zini et al., 2012).

The source of death certificates had the highest density among others, while medical records of hospitals and pathology reports had respectively the next grades. Although, regarding the coverage of the reports, medical records of hospital is in a lower level compared with the other two sources, it had an acceptable source density.

Death certificates and medical records of hospitals had the highest and lowest completeness. Generally, since patients with esophageal cancer, as a malignant disease, refer to diagnostic and therapeutic centers while having a poor survival (Mathers et al, 2001; Samadi et al, 2007; Boyle and Levin, 2008) the probability of being reported by each of the three sources is expected to be near each other and higher than the current report.

For the first time, this method was used for the sources used in cancer-registry systems and in order to make it possible to compare death certificates with the two other sources, the data of esophageal cancer which is a fatal cancer with a poor survival was applied. Besides, to compare the density between the fields and the sources, the empty or complete fields were considered not the accuracy and precision of the fields. In order to check the accuracy of filling the fields we need to apply more precise and efficacious methods (Scannapieco and Patini, 2004). Having a standard source or conducting a purposeful study can help to evaluate the accuracy of information in each data.

In fact controlling the quality of the registered data should be implemented for each registry system. While because of needing additional cost and time such process is not performed in many registries, using the presented method in this study we will be able to provide a general information about coverage, density and completeness when evaluating quality of data sources used in cancer-registry.

## References

- Boyle P, Levin B (2008). World cancer report 2008, IARC Press, International Agency for Research on Cancer.
- Bullard J, Coleman MP, Robinson D, et al (2000). Completeness of cancer registration: A new method for routine use. *Bri J Cancer*, **82**, 1111.
- Chen VW, Ferlay J, Padras JG, et al (1994). Comparability and quality control in cancer registration. International Agency for Research on Cancer(IARC).
- Elbasmi AA, Fayaz MS, AL-Mohanadi S, AL-Nesf Y, AL-Awadi A (2010). Reliability of the Kuwait Cancer Registry: A Comparison between Breast Cancer Data Collected by Clinical Oncologists and Registry Staff. *Asian Pac J Cancer Prev*, **11**, 735-8.
- Goldberg J, Gelfand HM, Levy PS (1980). Registry evaluation methods: A review and case study. *Epidemiol Reviews*, **2**, 210-20.
- Habibi A (1985). Epidemiological aspects of cancer in iran. *Int Surg*, **70**, 105.
- Hearst N, Hulley SB (1988). Using secondary data. Designing

- Clinical Research, Williams & Wilkins, Baltimore, 53-62.
- Kmet J, Mahboubi E (1972). Esophageal cancer in the caspian littoral of iran: Initial studies. *Science*, **175**, 846-53.
- Mathers CD, Boschi-Pinto C, Lopez AD, Murray CJL (2001). Cancer incidence, mortality and survival by site for 14 regions of the world. Geneva: World Health Organization: 8.
- Mohagheghi MA, Mosavi-Jarrahi A, Malekzadeh R, Parkin M (2009). Cancer incidence in tehran metropolis: The first report from the tehran population-based cancer registry, 1998 – 2001. *Arch Iranian Med*, **12**, 15-23.
- Mohagheghi MA, Mosavi-Jarrahi A (2010). Review of Cancer Registration and Cancer Data in Iran, a Historical Prospect. *Asian Pac J Cancer Prev*, **11**, 1155-7.
- Naumann F, Freytag J. C, Leser U (2000). Completeness of information sources, Professoren des Inst. Informatik. 19-23. Available from: <https://www.fing.edu.uy/inco/cursos/caldatos/articulos/Completeness.pdf>
- Naumann F, Freytag JC, Leser U (2004). Completeness of integrated information sources. *Information Systems*, **29**, 583-615.
- Roos LL, Roos NP, Fisher ES, Bubolz TA (1990). Strengths and weaknesses of health insurance. data systems for assessing outcomes. *Med Innovation at the Crossroads*, **1**, 47-67.
- Samadi F, Babaei M, Yazdanbod A, et al (2007). Survival rate of gastric and esophageal cancers in ardabil province, north-west of iran. *Arch Iran Med*, **10**, 32-7.
- Scannapieco M, Batini C (2004). Completeness in the relational model: a comprehensive framework. Available from: [Http://www.Dis.Uniroma1.It/~monsca/researchactivity/articoli/iq2004.Pdf](http://www.Dis.Uniroma1.It/~monsca/researchactivity/articoli/iq2004.Pdf).
- Sorensen HT (1992). Re: A compendium of public health data sources. *Am J Epidemiol*, **135**, 325-6.
- Sorensen HT, Ejlersen E, Moller-Petersen J, et al (1993). One-year follow-up after the first prescription of strong analgesics outside hospital. *Int J Risk and Safety in Med*, **4**, 209-14.
- Stone DH (1986). A method for the validation of data in a register. *Public Health*, **100**, 316-24.
- Zendehdel K, Hassanloo ZH, Sedighi Z, et al (2010). Improving quality of cancer registration in iran. Part2: Suggesting new national cancer registration program in iran, based on comparative study of 18 countries. *Hakim*, **12**, 50-7.
- Zini A, Nasser N, Vered Y (2012). Oral and pharyngeal cancer among the Arab population in Israel from 1970 to 2006. *Asian Pac J Cancer Prev*, **13**, 585-9.