

Estimating Completeness of Cancer Registration

—

An application to gastric cancer in the North Region Cancer Registry of Portugal

por

Clara Sofia Faria Cardoso de Castro

Tese de Mestrado em Métodos Quantitativos em Economia e Gestão

Orientada por: Professor Doutor Pedro Campos

Faculdade de Economia da Universidade do Porto

2010

Biography

Clara Sofia Faria Cardoso de Castro was born in Porto, Portugal, on September 3, 1987. In 2008, she graduated in Mathematics, minor in Statistics and Models, from the Science Faculty of the University of Porto, with a final grade of 14. In the same year, she was given a Scientific Initiation Grant from the Science and Technology Foundation (FCT) to develop a study regarding statistical analysis of extremes of environmental data. Also since 2008, she has been a student in the Faculty of Economics of Porto, attending a master degree on Quantitative Methods in Economics and Management, for which the curricular part was concluded with grade 16.

Since June 2009, she has been working in the North Region Cancer Registry (RORENO) as a statistician, collaborating in projects regarding cancer incidence and survival.

Acknowledgments

I would like to thank my supervisor, Professor Pedro Campos, for his support and guidance, which were crucial throughout this work, and to Dr. Maria José Bento, for proposing this theme and encouraging its development.

I also wish to express my gratitude to Dr. Ana Elvira, from ARS-Norte, for providing the access to death certificates; to Dr. Fabio Montanaro, for making STATA commands available for the flow method's application and to Dr. David Robinson and Luís Antunes, for the valuable comments regarding the application of some of the methods.

I am deeply grateful to my parents, my sister and Ricardo, for always being there and for all the support and patience.

Abstract

Completeness of registration is one of the most important quality indicators in a cancer registry. In literature, several methods are described to evaluate completeness, which are divided in two categories: qualitative methods and quantitative methods.

In this study, completeness values obtained by different methods are compared, for the particular case of gastric cancers registered in the North Region Cancer Registry of Portugal (RORENO) between 2001 and 2006, of patients resident in the district of Porto at date of diagnosis.

The qualitative methods used were Mortality/Incidence ratios and the proportion of microscopically verified cases and, among quantitative methods, the ones applied were the capture-recapture, the death certificates and M:I ratios method, and the flow method.

No significant differences were found between the results obtained through qualitative methods, although the decrease observed in mortality/incidence ratios indicates that there is an improvement in gastric cancer treatment. Overall completeness obtained through quantitative methods was of approximately 83% for the real data scenario, and of approximately 95% for the best case scenario considered in this study. In the real scenario, completeness of gastric registration was higher for female patients than for male patients.

Resumo

A exaustividade de registo é um dos indicadores de qualidade mais importantes num registo oncológico. Na literatura são descritos vários métodos para a estimação da exaustividade, que são divididos em duas categorias: métodos qualitativos e métodos quantitativos.

Neste trabalho apresenta-se uma comparação de resultados obtidos por aplicação de diferentes métodos, para o caso particular de tumores gástricos registados no Registo Oncológico Regional do Norte (RORENO) no período de 2001-2006, de pacientes residentes no distrito do Porto à data de diagnóstico.

Os métodos qualitativos usados foram os rácios M:I e a proporção de casos verificados microscopicamente. Os métodos quantitativos aplicados foram o de captura-recaptura, o método dos certificados de óbito e rácios M:I, e o método do fluxo de informação.

Não foram encontradas diferenças significativas nos resultados de exaustividade obtidos pelos métodos qualitativos, embora o decréscimo observado nos rácios mortalidade/incidência indique que há uma melhoria no tratamento do cancro gástrico. Os métodos quantitativos devolveram uma exaustividade global de aproximadamente 83%, no cenário real, e de aproximadamente 95%, num cenário considerado o melhor cenário possível. No cenário real, a exaustividade dos pacientes do sexo feminino foi superior à obtida para o sexo masculino.

Contents

1	Introduction	1
2	Cancer Registration	4
2.1	What is cancer?	4
2.2	The importance of cancer registration	5
2.3	Evolution of cancer registration	7
2.4	Cancer Registration in Portugal	10
2.5	Quality indicators in cancer registration	15
3	Available Methods – an overview	18
3.1	Semi-quantitative methods	18
3.1.1	Historic data methods	18
3.1.2	Mortality:incidence (M:I) ratios	20
3.1.3	Number of sources/notifications per case	21
3.1.4	Histological verification of diagnosis	22
3.2	Independent case ascertainment	23
3.3	Capture-recapture methods	24
3.4	Death certificate methods (DC methods)	28
3.4.1	DC and M:I method	29
3.4.2	Flow Method	31
4	Data and Methods	34
4.1	Semi-quantitative methods	36
4.2	Quantitative methods	36
4.2.1	Capture-recapture method	37

4.2.2	DC and M:I method	37
4.2.3	Flow method	38
5	Results	42
5.1	M:I ratio	42
5.1.1	Comparative results over time	42
5.1.2	Comparative results with some european registries	44
5.2	Histological verification of diagnosis	47
5.2.1	Comparative results over time	47
5.2.2	Comparative results with some european registries	49
5.3	Capture-recapture method	51
5.3.1	First scenario	51
5.3.2	Second scenario	52
5.4	DC and M:I method	53
5.4.1	First scenario	53
5.4.2	Second scenario	54
5.5	Flow method	55
5.5.1	First scenario	55
5.5.2	Second scenario	59
6	Conclusions	61
	References	67

List of Tables

2.1	Geographical coverage in the nine volumes of <i>Cancer Incidence in Five Continents</i> (source: Curado <i>et al.</i> (2007))	10
3.1	Values of incidence rates (per million) for upper and lower deciles of childhood cancer	19
4.1	Condensed description of the files used in the flow method application. . .	39
5.1	Comparative results of M:I ratios over time, for both sexes	43
5.2	Comparative results of M:I ratios over time, for male patients	43
5.3	Comparative results of M:I ratios over time, for female patients	44
5.4	Comparative results of M:I ratios from different registries, for both sexes .	44
5.5	Comparative results of M:I ratios from different registries, for male patients	45
5.6	Comparative results of M:I ratios from different registries, for female patients	45
5.7	Comparative results of MV% over time, for both sexes	47
5.8	Comparative results of MV% over time, for male patients	48
5.9	Comparative results of MV% over time, for female patients	48
5.10	Comparative results of MV% from different registries, for both sexes . . .	49
5.11	Comparative results of MV% from different registries, for male patients .	50
5.12	Comparative results of MV% from different registries, for female patients	50
5.13	Completeness values and respective confidence intervals obtained by adjusting model M_0 , in the first scenario.	51
5.14	Completeness values and respective confidence intervals obtained by adjusting model M_t , in the first scenario.	52

5.15	Completeness values and respective confidence intervals obtained by adjusting model M_0 , in the second scenario.	52
5.16	Completeness values and respective confidence intervals obtained by adjusting model M_t , in the second scenario.	53
5.17	DCI cases and completeness values obtained in the first scenario.	54
5.18	DCI cases and completeness values obtained in the second scenario.	54
5.19	Completeness of registration using deaths registered in 2005	55
5.20	Completeness of registration according to sex, using deaths registered in 2005	56
5.21	Completeness of registration using deaths registered in 2003	57
5.22	Completeness of registration according to sex, using deaths registered in 2003	58
5.23	Completeness of registration using deaths registered in 2001	58
5.24	Completeness of registration according to sex, using deaths registered in 2001	59
5.25	Completeness of registration assuming that death certificates are received for every patient that died	59
5.26	Completeness of registration according to sex, assuming that death certificates are received for every patient that died	60

List of Figures

2.1	The North Region of Portugal, divided in five districts: Braga, Bragança, Porto, Viana do Castelo and Vila Real.	12
2.2	Death-related information flow in Portugal (source: author)	14
3.1	Capture-recapture methods using two sources of information (source: author).	25
3.2	Registration of new cases of cancer, using death certificates (source:author).	28
3.3	DCI as a measure of completeness (source: Parkin & Bray, 2009).	29
3.4	Relationships between categories of registered and unregistered cancer patients at time t after diagnosis (source: Bullard <i>et al.</i> , 2000).	32
5.1	Completeness plus proportion lost, using deaths registered in 2005	56
5.2	Completeness plus proportion lost, using deaths registered in 2003	57
5.3	Completeness plus proportion lost, using deaths registered in 2001	59
5.4	Completeness plus proportion lost, assuming that death certificates are received for every patient that died	60

Chapter 1

Introduction

Cancer is a leading cause of death worldwide, affecting people at all ages. For its characteristics, cancer demands a specific registry which allows an up-to-date, permanent and comparable knowledge of cases, in order to define strategies to fight against it. In Chapter 2 of the present report, the importance of cancer registration and its evolution over time are described.

A cancer registry is, above all, an information source and, as such, the quest for excellence must be a priority. A high level of cancer registry quality is essential for drawing correct conclusions about cancer incidence, survival, treatment effectiveness, patterns of care, and also cancer control.

The main quality indicators regarding cancer registration are comparability, completeness, accuracy and timeliness. Completeness of cancer registration is defined as the proportion of all incident cancer cases in a defined population which are included in the cancer registry database. It is important to routinely measure completeness of population-based cancer registries because systematic bias in case reporting lead to the calculation of misleading and erroneous rates of cancer in the target population.

The goal of this study is to evaluate completeness of registration in the North Regional Cancer Registry of Portugal (RORENO) for malignant gastric cancer cases, excluding lymphomas, of patients resident in the district of Porto at date of diagnosis. RORENO is a population-based cancer registry which records every patient diagnosed with cancer,

provided that the patient lives in the North Region of Portugal at the time of diagnosis. The results presented in this work are the first ones to be obtained through quantitative methods in the assessment of completeness of cancer registration in Portugal and, particularly, in the North Region.

In Portugal, cancer is the second leading cause of death and it has a profound impact on patients, relatives and society, being probably the most feared disease among the general population. Gastric cancer was chosen to be included in this study because it is, for both sexes, one of the most common and lethal cancers in Portugal, specially in the North Region, so it is of great interest to evaluate completeness of registration of such tumours.

There are several methods available to evaluate completeness of cancer registration, that may be classified into two different categories: the semi-quantitative methods, which merely provide an indication of the degree of completeness relative to other registries or over time; and the quantitative methods, which give a numerical evaluation of the extent to which all eligible cases have been registered. In Chapter 3, the available methods are described, emphasizing the advantages and disadvantages of each one.

In order to apply any of the available methods, some data must be accessed and treated properly. In this study, two databases were accessed: gastric cancer cases registered in RORENO between 2001 and 2006 and death certificates collected regarding the same period of time. The attainment of the databases used and the modifications made to each one are described in Chapter 4.

In this study the semi-quantitative methods used were M:I (mortality/incidence) ratios and the proportion of microscopically verified cases, for which comparisons over time and with some other european countries were made. No significant differences were found between the values obtained by each of these methods, either over time or by comparison with other european registries. Nevertheless, the M:I ratios obtained decreased over time, which may indicate an improvement in health care related to gastric cancer treatment in the North Region of Portugal.

Among quantitative methods, the ones applied were the capture-recapture, the death certificates and M:I ratios method, and the flow method. When applying these methods,

two different scenarios were approached: one using the real data collected, and another which may be considered the best case scenario. The overall estimates for completeness of gastric registration, using the different methods, gave approximately the same value of 83%, rising to a value of approximately 95%, when using the best case scenario. When using the real data scenario, female patients presented higher values for completeness of registration (around 85%) than male patients (around 81%), for all methods.

Further detailed results obtained for each of the methods mentioned above, for all scenarios considered, are presented in Chapter 5 of this report.

Finally, the overall completeness values obtained are discussed in Chapter 6, where the limitations for each method are accessed and some future work is proposed.

Chapter 2

Cancer Registration

Cancer affects people at all ages, with the risk for most types increasing with age. According to the World Health Organization (WHO, 2009), the disease accounted for 7.4 million deaths (around 13% of all deaths) in 2004, being a leading cause of death worldwide.

In Portugal, cancer is the second leading cause of death and it has a profound impact on patients, relatives and society, being probably the most feared disease among the general population.

In this chapter, the importance of cancer registration and its evolution over time is accessed. The information flow in Portugal is also described, and the main quality indicators regarding cancer registration are addressed.

2.1 What is cancer?

A neoplasm (or tumor) is an abnormal mass of tissue that results when cells divide more than they should or do not die when they should. Neoplasms are quite heterogeneous, but may be divided in two groups: *benign neoplasms* and *malignant neoplasms* (also known as *cancer*).

Benign neoplasms usually do not imply that the patient's life is at risk. On the contrary, malignant neoplasms are potentially fatal since they are capable of invading and destroying the area of origin, and the adjacent areas. Also, malignant neoplasms can sometimes spread to other locations in the body via lymph or blood, a process referred to

as metastasis.

Depending on the cell and the tissue of origin, there are several types of cancer, for which clinical behaviour and treatment are very different. For that reason, Cancer cannot be faced as a single disease, but instead as multiple diseases.

The DNA changes which induce Cancer may be due to multiple causes: exposure to environmental agents (tobacco, excessive solar exposure), inadequate diet, persistent infection by some biological agents (bacteria or virus) or inheritance of genetic changes, causing a hereditary cancer predisposition (some colon and breast cancers).

2.2 The importance of cancer registration

For its characteristics, cancer demands a specific registry which allows an up-to-date, permanent and comparable knowledge of cases, in order to define strategies to fight it.

Globally, the role of a cancer registry is to contribute to the evaluation and control of cancer cases in a given population. Thus, a cancer registry must collect comprehensive information on all new cases of cancer occurring in that population, store that material securely and permanently, and analyse the information to allow for the production of regular reports.

In order to evaluate the burden of cancer in the world, data must be collected. When data regarding clinical description of the disease, identification of the patient, tumour, hospital and relevant physicians is put together with the information on treatment and follow-ups, relapses, metastases and, if applicable, date and cause of death, an invaluable database is created (Ruiz & Facio, 2004). This is the data needed to develop areas such as aetiological and epidemiological investigation, care planning and prevention, favouring both patients and society.

There are two main types of cancer registry: hospital-based and population-based.

The goal of a hospital-based cancer registry is to serve the needs of the host institution administration, the hospital's cancer programme and the individual patient. Since the design of a hospital-based registry is towards administrative and patient purposes, some of the data variables collected by hospital registries will be different from the ones obtained by a population-based registry. However, since many hospital-based registries submit

their data to a central population-based registry, they often include data variables needed by the central registry which have no utility for the hospital registry (Young, 1991).

A population-based cancer registry aims to collect the data which will give an accurate evaluation of cancer in a population, in order to understand and control the impact of cancer in that population. Analysis of the data collected provide the frequency of cancer cases and show which types are the most common. This will permit studies to identify the causes of cancer and, simultaneously, the registry data can be used to measure the significance of screening programmes or other procedures designed to reduce cancer incidence in the population. Registry data can also be used to study the importance of early diagnosis and of treatment (Esteban *et al.*, 1995). In this work, the type of cancer registry under consideration is a population-based one.

In this work, the cancer registry under consideration is a population-based cancer registry.

The role of a population-based cancer registry is to collect information on every cancer case occurring in a defined population (ENCR, 2003). In order to accomplish this task, a clearly delimited geographical territory must be fixed, and accurate data on the population living in that area obtained. Every effort must be made to exclude patients who do not live within the boundaries of the assigned area from the incidence data. It is also crucial to cover every potential source of data for residents of that territory (Esteban *et al.*, 1995).

Among the many uses of the data, one should emphasize

- **prevention** through the investigation to find the causes of cancer by looking for explanations for disparities in risk for different groups, and using material from the registry in education and information programmes;
- **detection** by evaluating the effectiveness of screening programmes and by identifying groups of people at greater risk who should be more frequently checked for cancer;
- **treatment** by determining whether every patient has the same access to treatment, and checking the effectiveness of different types of treatment.

Hence, no matter what type is the cancer registry, it is always useful when planning requirements for the personnel, medical facilities and equipment needed for the diagnosis

and treatment of the cancer patient. Any other method of counting and describing cancer will miss some cases and so cannot reliably tell if certain cancers are becoming more or less frequent, if the risk of developing cancer is different for people living in different areas or working in different jobs and even if cancer screening programmes are really producing an effect.

2.3 Evolution of cancer registration

The core activity of a population-based cancer registry is to produce statistics on the incidence of cancer, describing cancer patterns and trends. Although this was the focus of the registries to be launched in Europe and North America in the first half of the twentieth century, their role in providing information on other features of cancer occurrence and on the control of the disease has expanded progressively (Parkin, 2006).

Population-based cancer registration with an epidemiological focus began in the USA in 1935, when a division of cancer research was established in the Connecticut State Department of Health (Wagner, 1991). The Connecticut Tumor Registry started functioning on a statewide basis in 1941, registering cases from back to 1935. Further cancer registries were created in the USA and Canada in the early 1940s.

Probably the most significant stimulus for the worldwide establishment of cancer registries came from a conference that took place in Copenhagen in 1946 upon the initiative of Dr Clemmesen, Director of the Danish Cancer Registry. A group of 12 globally prominent specialists in the field of cancer control recommended the worldwide establishment of cancer registries to the Interim Commission for the World Health Organization (Clemmesen, 1974). Four years later, the World Health Organization established a Subcommittee on the Registration of Cases of Cancer and their statistical presentations which provided the first recommendations for the creation of cancer registries. Within a few years, European Nordic countries, Great Britain and Canada launched programs aiming to national coverage of registration (Morabia, 2004).

Another landmark was the International Symposium on Geographical Pathology and Demography of Cancer, organized by the *International Union Against Cancer (UICC)* in 1950, where the need for the specification of all new cases of cancers in a defined geographical area was stressed. Based on the recommendations of this Symposium, UICC established a *Committee on Geographical Pathology*. In 1965, the *International Agency for Research on Cancer (IARC)* was established as a specialized cancer research centre of the World Health Organization.

As a natural consequence of this development, the *International Association of Cancer Registries (IACR)* was formed in 1966 in Tokyo. The IACR serves as a membership organization for population-based cancer registries worldwide, and aims to establish standards for cancer registration, publish registry data and hold scientific meetings, thus collaborating closely with the IARC.

After the establishment of these organizations, many followed.

The *Surveillance, Epidemiology and End Results (SEER)* first began collecting data on cancer cases on January 1, 1973. This programme is a commanding source of information regarding cancer incidence and survival in the USA. It collects and publishes cancer incidence and survival data from different population-based cancer registries, the oldest of which is the Connecticut Tumour Registry, mentioned above.

Since its genesis in 1988, the *National Cancer Database (NCDB)* has been a cofunded project of the American College of Surgeons and the American Society of Cancer, having provided significant information based on data from over 25 million cancer patients diagnosed and treated in the USA (Winchester *et al.*, 2010).

In 1992 the Congress of USA attested the need of collecting more thorough data regarding cancer in the USA, and so the *National Program of Cancer Registries (NPCR)* was created, through the Cancer Registries Amendment Act. Another important organization connected to various institutions interested in improving the quality and use of data from registries is the *North American Association of Central Cancer Registries (NAACCR)*.

In Europe, the *European Network of Cancer Registries (ENCR)* promotes collaboration between cancer registries, defines data collection standards, provides training for

cancer registry personnel and regularly disseminates information on incidence and mortality from cancer in the European Union and Europe. This project has been in operation since 1990 and was established within within the framework of the *Europe Against Cancer Programme* of the European Commission.

Among european countries, Denmark must be mentioned for its long tradition in cancer registration. The Danish Cancer Registry was created in 1942, and it was the world's first program to register every cases of cancer arising in an entire nation (Jensen *et al.*, 1985).

The registries which are considered pioneers in Europe usually cover the whole of a country. Reporting of cancer cases is mandatory in some countries such as Finland, Poland, Norway and Slovakia, which have very disperse results depending on the resources available in each country. In Germany, a law was approved by the Federal Parliament that dictated the establishment of cancer registries in each of the 16 regions. In Spain, several population-based registries exist, each covering a different region: Albacete (created in 1990), Asturias (created in 1978), Basque Country (created in 1986), Granada (created in 1985), Mallorca (created in 1982), Murcia (created in 1981), Navarra (created in 1970), Tarragona (created in 1980), Zaragoza (created in 1960) and Girona (created in 1994). Nevertheless, other countries such as Greece have lacked a population-based for a long time.

In the last 20 years, the purpose of registries has extended further to incorporate the planning and evaluation of cancer control activities, and the care of individual cancer patients. Thus, population-based cancer registries are becoming more extensively implicated in studies of the process of clinical treatment of cancer patients.

In 1966, 32 registries published their results in volume I of *Cancer Incidence in Five Continents* (Doll *et al.*, 1966); in volume IX, forty years later, data from over 200 registries were included. The evolution in geographical coverage in the nine successive volumes of *Cancer Incidence in Five Continents* is provided in Table 2.1.

Volume	Year of publication	Registries	Populations	Countries	Period (approx)
I	1966	32	35	29	1960-1962
II	1970	47	58	24	1963-1967
III	1976	61	79	29	1968-1972
IV	1982	79	103	32	1973-1977
V	1987	105	137	36	1978-1982
VI	1992	138	166	49	1983-1987
VII	1997	150	183	50	1988-1992
VIII	2002	186	214	57	1993-1997
IX	2007	225	300	60	1998-2002

Table 2.1: Geographical coverage in the nine volumes of *Cancer Incidence in Five Continents* (source: Curado *et al.* (2007))

2.4 Cancer Registration in Portugal

In Portugal, studying cancer became a subject of interest at the end of the 19th Century. A landmark in such interest was an inaugural dissertation at the Medical-Surgical School of Lisbon, in 1889, denominated “The Carcinoma Microbe”, presented by Professor Câmara Pestana (ROR-Sul, 2008).

In 1906, a directive from the Ministry of the Realm nominated Professor Azevedo Neves as the secretary of a commission responsible for cancer study. This commission elaborated a survey which was sent to every Portuguese physician and for which results were published and successfully presented in Brussels, in 1908.

In 1963, the Statistics Department of the Instituto Português de Oncologia (IPO) started registering cancer cases treated at the institution. Though this data was conditioned to the pathologies treated at that health centre, it became published annually by the National Statistics Office. It was only in 1978 that a true hospital-based cancer registry was established at IPO.

The Cancer Registry of Viana do Castelo was the first population-based cancer registry of Portugal, initiating its activity in the early 1970s. Its establishment was due to the willing of Professor Daniel Serrão and Dr. José Maria de Carvalho and to the collaboration of most medical practitioners of the district of Viana do Castelo. In 1980 the

Registry of Viana do Castelo organized, with great success, the first meeting of GRELL (Group of Latin Language Registries) in Portugal, where three studies were presented using data collected by this registry. Although the interest for the incidence data was shared by the IARC authorities, and many studies were proposed to the registry, these were not fulfilled since it ceased operations in 1985, due to lack of resources.

The Cancer Registry of Vila Nova de Gaia began operating in 1981, and it was the first cancer registry which has published its data on a regular basis in the *Cancer Incidence in Five Continents* publications, since 1985. This fact evidences the quality of data collection, exploration and presentation in this registry.

The Portuguese Government, aware of the need to provide the country with such an important research tool, established in 1988, under the directive 35/88 of January 16th, that three regional cancer registries were to be created which, together, would comprise all of Portugal's geographical area: RORENO covering the North Region, ROR-Centro covering the Centre Region and ROR-Sul covering the South Region and Madeira islands. This situation is maintained to the present, and the Regional Cancer Registries embrace every cancer registry from health centres and hospitals of the respective region.

Cancer is the second leading cause of death in Portugal and it has a profound impact on patients, relatives and society, being probably the most feared disease among the general population. The recognition of these facts led to the fact that fighting against cancer is nowadays one of the main priorities in the National Health Plan. The creation of a National Plan for Prevention and Control of Cancer (PNPCDO) reflects the need to establish a comprehensive strategy for action in different areas related to cancer prevention and treatment, in order to achieve higher quality and equity of care.

The most important goals of the PNPCDO 2007-2010 are to reduce morbidity and mortality from cancer, and to improve quality of life and patient satisfaction with health care provided. On that account, it is important to:

- Promote healthy lifestyles especially through obesity prevention and smoking combat;
- Define good practices in diagnosis and treatment and improve the accessibility and equity in care provision;

- Grant the access to paliative care and psychosocial support to patients and their relatives;
- Encourage scientific research, expecting for its contribution to improving quality of care;
- Support training to ensure te existence of human resources, necessary to imple-ment the plan;
- Monitor and evaluate the measures implemented and results.

Among the three regional cancer registries available in Portugal, this study will focus on the North Region Cancer Registry.

The Registo Oncológico Regional do Norte (RORENO) registers all new cancer cases in the North Region of Portugal, as mentioned above. According to the latest census, in 2001, the population covered was of 3 236 089 inhabitants distributed in five districts (Braga, Bragança, Porto, Viana do Castelo and Vila Real), which represented about 30% of the portuguese population. The map of the North Region of Portugal is shown in Figure 2.1.



Figure 2.1: The North Region of Portugal, divided in five districts: Braga, Bragança, Porto, Viana do Castelo and Vila Real.

The registry is placed at the Instituto Português de Oncologia, in the district of Porto, which is the leading cancer treatment centre of the north region, granting radiotherapy, surgery and chemotherapy to all patients. Two general hospitals are also available in that

district which are qualified to provide cancer treatment. However, some kinds of cancer may be treated in the other districts. The health system is based on a network of public primary health centres and hospitals, in conjunction with private clinics and laboratories. Screening programmes are mostly opportunistic and include cancers of the breast, cervix and colorectum. In Bragança, a population-based screening program for breast cancer was launched in 2001, comprising 22 541 women.

The main procedure for gathering data is passive notification of cases. In 2000, 81.4% of registrations were collected from 32 public health institutions and 18.6% from eight private pathology laboratories and hospitals (Curado *et al.*, 2007). The registry's staff is responsible for treating the incoming information from the health institutions, evaluating the quality of that data and fulfilling the data requests for epidemiological studies. Nowadays, the registry is staffed by a medical epidemiologist, a full-time registrar and two full-time and one part-time statisticians. Two medical pathologists are also involved as consultants.

Every year since 1988 an annual report on cancer incidence in the North Region of Portugal has been published. Plus, registry data has been used to carry some projects of the Regional Health Administration, namely in provision of screening services.

Although the registry was established on a legal framework, the access to a unique personal identification number is denied. This is due to the data confidentiality, an issue which is extremely important in the operation and maintenance of every cancer registry. Given their need to access individual and detailed information, cancer registries have always functioned under rigorous conditions of respect for the confidentiality of medical information, specially in the security of data files and the release of information to third parties.

Current developments in biological sciences have promoted a growing concern in creating even more restricted ethical guidelines, leading to the need for a signed informed consent for its collection, storage and usage. Even though this was motivated by attending for individual rights, it frequently collides with social responsibilities, as are disease notification and posterior registration.

Cancer registration is impracticable if individual consent is demanded from cancer patients (Ingelfinger & Drazen, 2004). A persistent campaign carried by epidemiologists and public-health experts depreciated the informed-consent principle for data collected under public-health intentions (for example, in the USA, Japan and the European Union).

Aiming to the privacy protection of every oncologic patient and to insure that the supplied informations are not abusively used, RORENO follows the confidentiality guidelines provided by the International Agency for Research of Cancer – IARC (2004) and the European Network of Cancer Registries – ENCR (2003).

However, some local authorities that produce ethical guidelines still require, as it is mentioned in the Declaration of Hensinki (2008), that “in medical research involving competent human subjects, each potential subject (...) must be informed of the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal.”

Another subject worth describing is the process of collecting death-related information. In Portugal, when a person dies, the conservatory of that person’s birth place receives a death notification. This information is passed on to the health center to which the person belongs and, thereafter, to one of the Health Sub-Regions, which may hand the information to the respective regional cancer registry. The figure 2.2 summarizes this process.

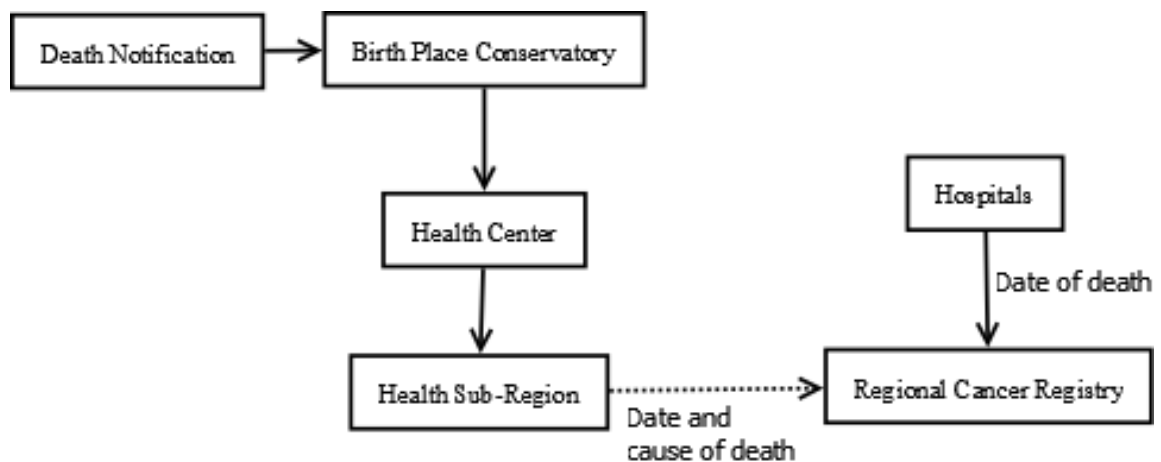


Figure 2.2: Death-related information flow in Portugal (source: author)

In the Northern Health Sub-Region, that information is received in paper, and it is not computerized afterwards. In 2005 and 2006, a major effort was undertaken by the Northern Health Sub-Region, and the information regarding patients who had a mention of cancer in the death certificate was sent to RORENO. Previously, only hospitals and private clinics working in collaboration with RORENO provided information on the patients' date of death.

In 2008, RORENO has been granted the access to a continuously updated National Health Database, in which the registry searches every registered patient on a regular basis in order to know if the patient is still alive or, otherwise, to record the date of death.

All of these improvements have led to more up-to-date and complete data in the North Region Cancer Registry.

2.5 Quality indicators in cancer registration

A cancer registry is, above all, an information source and, as such, the quest for excellence must be a priority (Skeet, 1991).

Before it is appropriate to measure any aspect of the quality of cancer registry data, it must be noticed that the use of several sources of cancer notification, the establishment of data definitions and the permanent training of tumor registrars are some of the requirements to create a cancer registry of high quality (Schouten, 1996).

A high level of cancer registry quality is essential for drawing correct conclusions about cancer incidence, survival, treatment effectiveness, patterns of care, and also cancer control. In this section the main quality indicators regarding cancer registration are briefly addressed.

Comparability is the extent to which coding and classification procedures, together with the definitions of recording, agree with international guidelines.

The comparability of statistics produced for distinct populations, and over time, is crucial to their interpretation. Hence, the comparability of cancer data is evaluated through a review of the registration procedures, including a specification of the standards and definitions that have been attended to. In the assessment of comparability of cancer registry

data, four items require a special concern:

- The system used for classification and coding of neoplasms;
- What is the definition of an incident case and of an incident date;
- The difference between a primary case and an extension, recurrence or metastasis of an existing one;
- The recoding of cancers detected in asymptomatic individuals.

The **completeness** of cancer registry data is defined as the proportion of all incident cancer cases in a defined population which are included in the cancer registry database. In theory, all cases of cancer in a defined population are registered in a population-based cancer registry. It is important to routinely measure the completeness of population-based cancer registries because systematic bias in case reporting lead to the calculation of misleading and erroneous rates of cancer in the target population. Incidence rates and survival proportions will be close to their true value if maximum completeness in case-finding procedures can be achieved.

Within each institution, there are various routine casefinding procedures, such as review of the disease indices, pathology reports, radiation therapy logs, among others. Rigorous case completeness audits can be difficult, time consuming and costly, especially in the larger or more populated geographical areas. Nevertheless, since correct assumptions can only be made from studies based on valid and unbiased data, the difficulty and expense must be accepted as a necessary cost of registry operations (Goldberg *et al.*, 1980). In the next chapters, completeness of registration is the quality indicator focused.

Validity (or accuracy) is defined as the proportion of cases registered with a given characteristic (e.g., sex, age or diagnosis) which truly have this attribute. It depends on the veracity of source documents and the level of capability in abstracting, coding and recording this information in the registry data base.

There are four available methods for evaluating validity of cancer registry data, which are described in detail elsewhere (Parkin *et al.*, 1994): *diagnostic criteria method* (histological verification and death certificate only); *missing information analysis*; *reabstract-*

ing and recoding; and *internal consistency method*. These methods provide numerical indices of validity, at least on an interval scale, and consequently allow for comparisons with other registries or, within a registry, over time, or with respect to specified subsets of cases.

Timeliness of reporting of cancer registry relates to the rapidity at which a registry collects, processes and reports satisfactorily dependable and complete cancer data. This clearly influences the extent to which data are complete and accurate (Bray & Parkin, 2009).

Expeditionous reporting of information on cancer cases is considered a priority for cancer registries. Swift access to cancer information is of clear advantage to both providers and researchers, and the early supply of data usually increases the registry's reputation. However, since registries are permanently updating their database as reports are received, and some notifications arrive long after the case was diagnosed, statistics for the recent periods will be incomplete, and in need for further updates. There is, therefore, some conflict between the requirement for timely data, and others aspects of data quality, namely completeness.

Chapter 3

Available Methods – an overview

There are several methods available to evaluate completeness of cancer registration. It is useful to separate those methods into two categories. While the qualitative (or semi-quantitative) methods merely provide an indication of the degree of completeness relative to other registries, or over time, the quantitative methods give a numerical evaluation of the extent to which all eligible cases have been registered (Parkin & Bray, 2009).

In this chapter, the available methods for evaluating completeness of registration are described, focusing in more detail the quantitative methods.

3.1 Semi-quantitative methods

A number of methods provide some indication of the completeness of a registry, but do not actually quantify the number of cases missing. The ones described below are some of the methods which belong to this category.

3.1.1 Historic data methods

The distribution of cases registered each year allow to check on the **stability of incidence rates over time**, and to notice potential problems in the registration process and/or in the population during the period under consideration. Unexpected or implausible trends in incidence, which cannot be explained by inconsistencies in the estimation of the population at risk (part of the population which is susceptible to develop a specific cancer), are

potential demonstrations of changes in completeness of registration.

This idea can be enlarged to include comparisons of results with the ones observed in other populations, that might have been expected to show resembling rates. Assuming that the incidence rates for specific cancers tend to be similar in datasets from the same region, region-specific standards can be defined for the expected incidence rates, so that observed values can be compared with those standards (Parkin & Bray, 2009). Differences from regional standards may reflect specific local oscillations in prevalence of risk factors, or the presence/intensity of screening. However, systematic disparities provide evidence of possible under-registration (or over-registration due, for example, to duplicate records).

Also the **age-specific incidence curves** should be examined, to detect unexpected variations in the anticipated patterns. These curves include, for a single time period, data from many birth cohorts. If there are changes in risk for specific cancers in different cohorts, these will be revealed in the shape of the curve of incidence with age, although such cohort effects can only be noticed when data are available from several time periods (Parkin *et al.*, 1994). Taking this into account, the shape of an age-specific incidence curve is an important indicator of possible under-ascertainment.

Regarding **childhood cancers**, the incidence rates in the childhood age groups (0-4, 5-9, 10-14) reflect much less variability than in adults. In this case, the possibility of under (or over) registration can be analysed by comparing the observed age-specific rates with an expected range of values. The values which lead to suspicion of under or over registration, published in Volume VIII of *Cancer Incidence in Five Continents* (Parkin & Plummer, 2002), are shown in Table 3.1.

Age	Boys		Girls	
	Lowest	Highest	Lowest	Highest
0 – 4	< 12.3	> 24.7	< 9.7	> 21.4
5 – 9	< 8.5	> 15.6	< 6.9	> 12.0
10 – 14	< 8.5	> 15.0	< 6.8	> 13.6

Table 3.1: Values of incidence rates (per million) for upper and lower deciles of childhood cancer

So, for instance, if the incidence rate obtained for boys in the age group 0–4 is 11.7

per million, one should suspect of under-registration on that specific type of cancer. If it is 24.8 per million, one should suspect of over registration.

3.1.2 Mortality:incidence (M:I) ratios

M:I ratios are comparisons of the number of deaths attributed to a specific type of cancer and the number of incident cases in the same period of time.

When the quality of the mortality data is good, the M:I ratio is close to the case fatality, measured as 1 minus 5 year relative survival probability. The probability of survival is evaluated in intervals of time elapsed since diagnosis (1, 3, 5-year survival). The *observed survival* proportion is influenced by mortality both from the cancer of interest and from other causes. For this reason, *relative survival* is usually estimated, which is defined as the ratio of the observed survival in the group of patients to the survival expected in a group of people in the general population without the disease of interest.

On the other hand, when mortality statistics are of poorer quality, the relationship will be less close, so the evaluation of the M:I ratio should take this into account. In particular, if the M:I ratio exceeds 1, it is usually a signal of under-registration due, for example, to innacurate cause of death or incomplete information from the death certificate (Parkin *et al.*, 1994).

To compare observed M:I ratios with region-specific standard values, checking on significant differences, the test described below may be used (Parkin & Bray, 2009).

Let us define:

- Comparison populations are registries from the same region, or from regions indexed by $i = 1, \dots, n$, where n is the total number of registries;
- d_i = number of incident cases in registry i ;
- m_i = number of deaths in registry i .

We start with a *Poisson* model for m_i/d_i in which the ratio of expected values is θ . This model is converted to a binomial model by conditioning on the total number of cases and deaths $n_i = m_i + d_i$. Then, the estimator for the ratio of expected values for all registries, $\hat{\theta}$, is given by

$$\hat{\theta} = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n d_i} \quad (3.1)$$

Let us define

$$\hat{\phi} = \frac{1}{n-1} \sum_{i=1}^n \frac{(m_i - \hat{\theta}d_i)^2}{n_i \hat{\theta}} \quad (3.2)$$

For the registry under test, with cases d_j and deaths m_j , the test is formulated as:

$$H_0 : \frac{m_j}{d_j} = \theta \quad (3.3)$$

$$H_1 : \frac{m_j}{d_j} \neq \theta$$

for which the test statistic is given by

$$Z^2 = \frac{(m_j - \hat{\theta}d_j)^2}{\hat{\phi}n_j \hat{\theta}} \sim \chi^2_{(1)} \quad (3.4)$$

So, consulting the chi-square table, using a 5% significance level, the null hipotesis is rejected if $Z^2 \geq 3.84$ and, in this case, the registry is flagged as unusual.

3.1.3 Number of sources/notifications per case

Most population-based cancer registries use several routine case-finding procedures, such as review of pathology and autopsy reports, hospital patient-disease information systems, radiotherapy notes and death certificates. Using as many sources as possible reduces the possibility of cancer diagnoses going unreported. Thus, completeness of registration is increased and the quality of the data is enhanced by bringing together every information relating to each patient in a single file.

Two indices have been used as indicators of completeness: the average number of sources per case, and the average number of notifications per case. The higher the average number of sources per case, the higher is likely to be the degree of completeness in the registration process. In considering notifications from a single source, those arising from the same episode should not be identified as being separate notifications of the same case.

3.1.4 Histological verification of diagnosis

The indicator ‘Percentage of cases morphologically verified’ (MV%), given by:

$$MV\% = \frac{\text{Number of cases diagnosed via microscopic verification}}{\text{Total number of diagnosed cases}} \quad (3.5)$$

is a measure of the validity of the information in a registry, and methods for comparing observed and expected values of MV% were described by Bray & Parkin (2009). However, a very high proportion of cases diagnosed by histology or cytology/haematology (higher than reasonably expected) suggests over-reliance on the pathology laboratories as a source of information, and failure to find cases diagnosed by other means.

The percentage of cancer cases likely to be histologically verified for a given site is dependent upon local circumstances. It might be low if the means for taking biopsies, or examining the tissue, are lacking or inadequate. Conversely, the availability of sophisticated imaging techniques may reduce the need for biopsy. Tables of average values for MV% by site and world area are given in Volume IX of *Cancer Incidence in Five Continents* (Shin *et al.*, 2007), as a guide to what is expected in a particular registry.

To compare the registry MV% by site with values from registries in the same region, or from regions indexed by $i = 1, \dots, n$, checking on significant differences, the test described below may be used (Parkin & Plummer, 2002).

We start by defining:

- y_i = number of microscopically verified cases in registry i ;
- d_i = total number of cases in registry i .

Using the binomial model for y_i :

$$E(y_i) = pd_i \quad (3.6)$$

$$Var(y_i) = p(1 - p)d_i$$

where p is the proportion of MV cases.

The overdispersion model changes this to

$$Var(y_i) = \phi p(1 - p)d_i \quad (3.7)$$

The parameters p and ϕ are estimated by

$$\hat{p} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n d_i} \quad (3.8)$$

and

$$\hat{\phi} = \frac{1}{n-1} \sum_{i=1}^n \frac{(y_i - p d_i)^2}{d_i p (1-p)} \quad (3.9)$$

For the registry under test, with data d_j and y_j , the test is formulated as:

$$H_0 : \frac{y_j}{d_j} = p \quad (3.10)$$

$$H_1 : \frac{y_j}{d_j} \neq p$$

for which the test statistic is given by

$$Z^2 = \frac{(y_j - \hat{p} d_j)^2}{\hat{\phi} \hat{p} (1 - \hat{p}) d_j} \sim \chi_{(1)}^2 \quad (3.11)$$

So, consulting the chi-square table, using a 5% significance level, the null hipotesis is rejected if $Z^2 \geq 3.84$ and, in this case, the registry is flagged as unusual.

3.2 Independent case ascertainment

There are two main procedures available to evaluate completeness through independent case ascertainment: **re-screening of cases** and **using independent sources of cancer cases**.

Re-screening of cases relates to the approach whereby one of the sources of a cancer registry is the subject of a ‘case-finding audit’. This audit is a re-examination of that source’s materials from a time period for which the registry has finished its case-ascertainment. The source may be selected either randomly or because of apparent under-reporting. The results of this independent inspection of records are then linked to the routine registrations. After the verification of reportability of unmatched cases, the true percentage of cases missed that should have been registered is calculated, representing the rate of under-registration.

Such studies usually focus on only one type of source (mainly, hospital reporting), providing an estimate for the completeness of reportable cases by individual facilities, instead of incident cancers (Parkin *et al.*, 1994). Thus, it is not a true measure of completeness for a multi-source population-based registry, and should be interpreted with caution. An application of this method is provided, for example, by Tingulstad *et al.* (2002).

The use of one or more data bases containing cancer cases, obtained *independently* from the cancer registry, is a particularly valuable and direct method of measuring completeness (Parkin & Bray, 2009). It requires matching the independent case series to the cancer registry database, in order to flag the cases missed by the registry. The proportion of suitable cases who are already registered is an objective quantitative measure of completeness. The independent sources more usually found in literature are patients involved in multi-centre clinical trials or multi-hospital case-control studies, cases identified in cohort studies or by community surveys, and patients recorded in hospital databases which are not accessed by the registry.

As cancer registries usually attempt to use all possible sources of information to target the cancer cases in the covered populations, it is not always possible to apply this method. However, it has been widely used, either to evaluate completeness of registration for all cancers combined (Villard-Mackintosh *et al.*, 1988; Parkin *et al.*, 2001), or for a single site (Brewster & Stockton, 2008).

3.3 Capture-recapture methods

Capture-recapture methods were originally conceived as a tool to estimate the size of free-living animal populations. Succintly, samples of animals are captured, tagged, released and then recaptured, and the size of the animal population is estimated from the numbers of animals captured and recaptured in each sample. Since cancer registries use multiple data sources, capture-recapture methods can be applied to these incomplete lists of patients to evaluate completeness of registration. Such sources may include notifications by clinicians, radiologists, pathologists or through death certificates (Brenner *et al.*, 1995). The simplest capture-recapture method includes two sources of information. The

idea behind this method is represented in Figure 3.1.

Casos identificados		na fonte A	
		Sim	Não
na fonte B	Sim	n11	n01
	Não	n10	n00 = ?

Figure 3.1: Capture-recapture methods using two sources of information (source: author).

The capture-recapture method assume that:

- there is no dependency in the capture process between all sources in a multi-source model;
- all individuals have the same probability of being captured.

In practice, it is likely that these assumptions do not hold in cancer registration: cases captured by one source may have higher, or lower, probability of being captured by the others (violating the first assumption); it is also possible that some individual characteristics are related to the probability of capture (violating the second assumption).

Another disadvantage of the capture-recapture method, as mentioned in Silcocks & Robinson (2004), is that while small sample sizes result in wide confidence limits, large sample sizes are inoperable because of the workload concerned. Although some cancer registries have applied the capture-recapture method successfully (Brenner *et al.*, 1995; Hay, 1997; Larsen *et al.*, 2009) and some previous studies encouraged a careful application of this method to supplement the traditional ones (Brenner *et al.*, 1994), all these problems have lead to distrust capture-recapture methods in the evaluation of completeness (Tilling, 2001). However, some ways to deal both the problem of dependency between sources and the problem of the characteristics associated with capture are available.

A simple method to allowing for the inequality of the capture probabilities is to stratify the analysis by the variables which may be considered related to the capture process and to apply the capture-recapture method to each stratum separately. It must be noted, however, that this increases the variability of the estimated population size for each stratum, particularly when data are sparse.

To deal with the problem of dependency between sources, Larsen *et al.* (2009) used Brenner's method (Brenner, 1995) to evaluate the dependence between pairs of sources and then grouped the most dependent sources, before estimating the missing cases in a two way method with a third source. A different approach, useful when sources are all dependent, is to handle dependence between sources through log-linear models (Bishop *et al.*, 1975). In this case, the best option is to use a model with all possible interactions between sources (Regal & Hook, 1991). To estimate the population size using a log-linear model, up to three sources of variation can be incorporated among capture probabilities: a temporal effect (subscript t), a heterogeneity between individuals (subscript h) and a behavioral effect (subscript b). These sources of variation lead to eight fundamental closed population models: M_0 (no source of variation), M_t , M_h , M_{th} , M_b , M_{tb} , M_{bh} , M_{tbh} , which are explained below.

Consider a closed population of N individuals undergoing a capture-recapture experiment with t capture occasions. The capture history of one individual can be expressed as a $t \times 1$ vector $w = (w_1, \dots, w_t)$ where $w_j = 1$ if the individual is captured at the j th occasion and 0 if not. Define n_w as the number of individuals in the population with capture history w . Then, the total number of individuals caught at least once is $n = \sum n_w$, taking into account the $s = 2^t - 1$ capture histories including at least one capture.

A statistical model has a predicted frequency μ_w associated to n_w . The goal is to estimate μ_0 , which is the number of individuals missed in the experiment. For the models mentioned above, μ_w is log-linear and it can be expressed in terms of an $s \times d$ design matrix X where d is the size of θ , the vector of log-linear parameters.

Model M_0

The model M_0 is the simplest model in which all capture occasions are independent with a common probability p of being caught. For this model, we have $\theta = (\gamma, \beta)'$, where $\gamma = \log\{N(1 - p)^t\}$ and $\beta = \log\{p/(1 - p)\}$ and the predicted value for the frequency of the missed units is $\mu_0 = N(1 - p)^t = e^\gamma$.

Models M_t , M_h and M_{th}

In model M_t , each capture occasion has its own capture probability, say p_j for occasion j . The design matrix has $d = t + 1$ columns $X_0 = 1$ and $X_j = w_j, j = 1, \dots, t$. It involves

intercept $\gamma = \log\{\prod(1 - p_j)\}$ and parameters $\beta = \log\{p_j/(1 - p_j)\}, j = 1, \dots, t$. The log-predicted frequency is

$$\log(\mu_w) = X_0\gamma + \sum_1^t X_j\beta_j. \quad (3.12)$$

Under this model, the predicted value for the frequency of missed individuals is, once again, $\mu_0 = e^\gamma$.

Heterogeneity in capture probabilities of the N individuals may be modelled by adding a constant interaction to (3.12) for all pairs of occasions, which leads to the log-linear model for M_{th} having $t + 2$ parameters:

$$\log(\mu_w) = X_0\gamma + \sum_1^t X_j\beta_j + X_{t+1}\alpha, \quad (3.13)$$

where $X_{t+1} = \sum_{j>k} w_j \times w_k$ is the column of X for heterogeneity.

Setting all the β_j equal to the same value β in (3.13) gives a log-linear model for M_h .

Models M_b and M_{bh}

When an individual's behaviour changes after the first capture, the capture probability goes from p to c . When estimating N , c is a nuisance parameter. A set of sufficient statistics for (N, p) is $\{u_j : j = 1, \dots, t\}$, where u_j is the number of individuals that are first caught at the j th occasion. If one lets $\mu_j = E(u_j)$, M_b is a log-linear model with

$$\log(\mu_j) = \gamma + (j - 1)\beta, \quad j = 1, \dots, t \quad (3.14)$$

where $\gamma = \log(Np)$ and $\beta = \log(1 - p)$. The size of the population is estimated by $N = e^\gamma/(1 - e^\beta)$ and the number of individuals missed is $\mu_0 = e^\gamma e^{\beta t}/(1 - e^\beta)$.

To also model heterogeneity in capture probabilities, one must assume that the first t_0 u -values do not follow model (3.14), since individuals caught early are not representative of the ones not caught after $t_0 \geq 0$ occasions. One has $\log(\mu_{j+t_0}) = \gamma + (j - 1)\beta$ for $j = 1, \dots, t - t_0$ where $\gamma = \log(N_1 p)$ and $\beta = \log(1 - p)$, in which N_1 stands for the size of the population not captured after t_0 occasions and p is the capture probability on capture occasions $t_0 + 1, \dots, t$. The size of the population N is then estimated by $N = \sum_1^{t_0} E(u_i) + e^\gamma/(1 - e^\beta)$.

Model M_{bt}

Model M_{bt} may be expressed as a log-linear model by imposing linear constraints on the parameters (Rivest & Lévesque, 2001). This model incorporates both a behavioural and a time effect.

A more recent alternative to handle dependence between sources is the inclusion of covariates in a logit model, which also deals with the problems of the capture-related characteristics (Tilling & Sterne, 1999).

In order to compare available capture-recapture methods, a simulation study was carried by Schmidtman (2008), in which all estimators tended to overestimate the total number of cases therefore leading to an underestimation of completeness.

3.4 Death certificate methods (DC methods)

Death certificates are one of the main sources of information on a cancer registry, and have three main uses in cancer registration (Parkin *et al.*, 1994): (i) as a complementary source of information on new cancer cases, (ii) for quality control for both completeness and validity and (iii) for studies on survival of registered patients. In this section we address the access to death certificates as a means of capturing information on cases that were not registered during life.

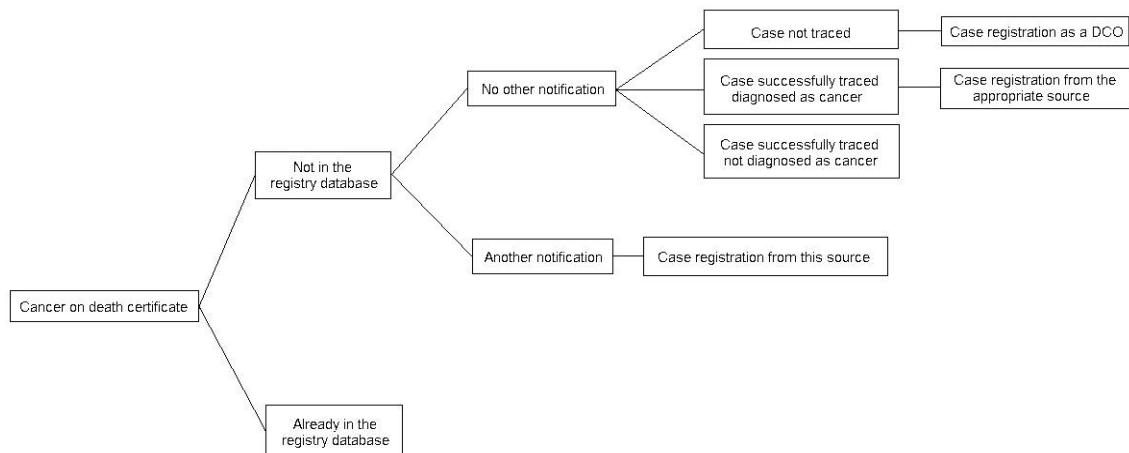


Figure 3.2: Registration of new cases of cancer, using death certificates (source:author).

A “Death Certificate Initiated” (DCI) registration is one for which any information available from other sources was found as a result of trace-back procedures, initiated because of a first information via death certificate. It is important to note that DCI cases are different from DCN (Death Certificate Notification) cases, the ones for which subsequent information was received *without* the need of a trace-back enquiry. DCI registrations also exclude cases subsequently found not to be cancers (Figure 3.2).

After all the trace-back procedures performed on DCN cases, the cases for which no other information than a death certificate mentioning cancer was obtained are termed DCO (Death Certificate Only) cases. Therefore, although the DCO% is not an indicator of completeness of registration, an elevated DCO% is suggestive of incompleteness.

Two methods of evaluating completeness based on death certificates are available, allowing the use of all cases or subsets defined by site or age group.

3.4.1 DC and M:I method

According to this method, the indicator of completeness is the proportion of cases first reported to the registry by other means than the death certificate. It requires that DCI cases can be explicitly identified by the registry, and it estimates the proportion of the initially unregistered cancer cases that do not die using the mortality:incidence ratio (M:I). The idea of this method is illustrated in Figure 3.3.

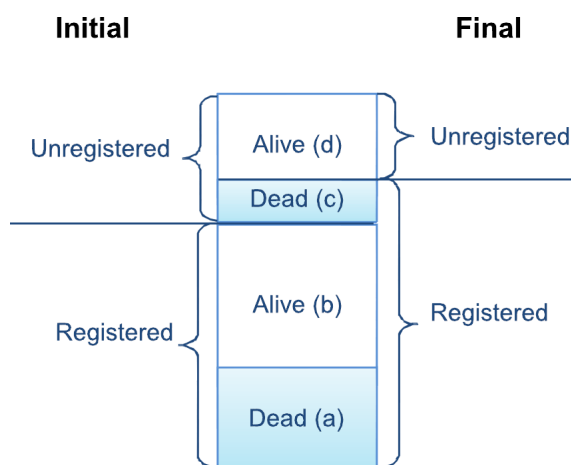


Figure 3.3: DCI as a measure of completeness (source: Parkin & Bray, 2009).

After the inclusion of DCI cases in the registry data base, the final degree of under-

registration is given by $d/(a + b + c + d)$. In order to obtain a numerical estimate of d , it is assumed that the proportion of unregistered cases which die ($c/(c + d)$) is equal to the proportion of registered cases which die ($a/(a + b)$). Thus, $d = bc/a$ represents the number of undetected cases still alive, so the degree of completeness may be estimated as:

$$\frac{\text{Final registrations}}{\text{Final registrations} + d} = \frac{a + b + c}{a + b + c + d} \quad (3.15)$$

The data required for this estimate are the proportion of cases registered because of information from a death certificate ($c/(a + b + c)$) and, for the cases registered during life, the proportion that die ($a/(a + b)$). An approximation to this case fatality is the mortality/incidence ratio (M:I) which is usually obtained from death registration, independently of individual follow-up of registered cases. Although the M:I ratio includes DCI cases ($(a + c)/(a + b + c)$), in practice this is not very different from the case fatality of non-death certificate cases ($a/(a + b)$), since the proportion of DCI cases is relatively small, say <10% (Parkin & Bray, 2009). If this occurs, the estimate for completeness can be based on DCI and the M:I ratio, both expressed as proportions, to yield (Parkin *et al.*, 1994):

$$\frac{1}{1 - DCI + DCI \times (1/M : I)} \quad (3.16)$$

Ajiki *et al.* (1998) modified this equation to apply to not so low proportions of DCI cases, also considering DCI and the M:I ratio expressed as proportions. This modification led to:

$$\frac{1 - DCI \times (1/M : I)}{1 - DCI} \quad (3.17)$$

The assumption behind this estimate is that the case fatality is alike for registered and unregistered cases, which is probably untrue, since other studies suggest that unregistered cases are likely to have higher fatality than cases detected by the usual case-finding procedures of the registry. Another assumption is that M:I ratios are relatively constant over the short time periods under consideration.

The derivation of both formulas is given in detail by Kamo *et al.* (2007).

3.4.2 Flow Method

Introduced by Bullard *et al.* (2000), this method is based on the logical flow of data in the registration system, and on the time distribution of various probabilities inherent to this flow, which can be calculated using routine cancer registry data.

Using information on survival of registered cases it estimates the cases not traced via death certificate and, based on information about cancer patients who have died, it derives the probability of registration at different time intervals post-diagnosis. It does not require re-abstraction of data, since the DCI cases don't have to be explicitly enumerated, and can be executed rapidly and inexpensively.

The flow method assumes that all death certificates mentioning malignant disease are received by the registry and that, after death, cancer patients are not registered from sources other than death certificates.

Three time-dependent probabilities are defined as follows:

- $s(t_i)$ is the probability that a cancer patient is still surviving at time t_i after diagnosis. This can be derived by a conventional survival analysis, using the actuarial method. If possible, DCOs should be included in the analysis as incident cases in the year they occur, being their survival imputed from that of the DCI patients whose details were successfully traced back from the death certificate and who were diagnosed in the required period.
- $m(t_i)$ is the probability that the death certificate of a patient who dies in the time interval (t_i, t_{i+1}) after diagnosis mentions cancer. This is obtained for the cancer patients who died in the survival analysis. The numerator is the number of deaths in the given interval since diagnosis for which the death certificate includes a mention of cancer, and the denominator is the total number of deaths in the same interval.
- $u(t_i)$ is the probability that a patient surviving until time t_i after diagnoses is still unregistered. This can be estimated from cancer registrations for patients who are now dead, being based on a survival model with registration as the event and censoring at death. This approach is refined by censoring one year before death, because the probability of registration among those who die of cancer increases during the year or so before death.

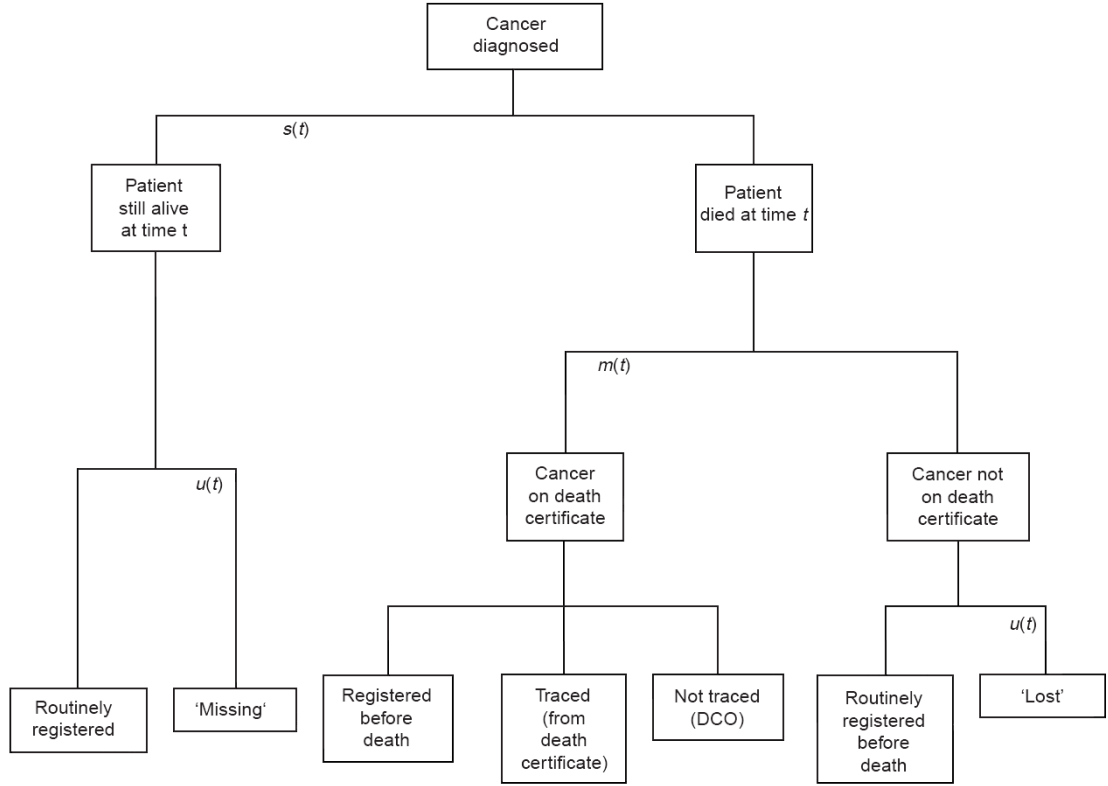


Figure 3.4: Relationships between categories of registered and unregistered cancer patients at time t after diagnosis (source: Bullard *et al.*, 2000).

Using the probabilities described above, we can derive the proportions of missing (M) and lost (L) cases, among the cases remaining unregistered after inclusion of DCI cases. Figure 3.4 clarifies the flow method procedures. The proportion of cases still alive and unregistered at time t_i after diagnosis ($M(t_i)$) is given by:

$$M(t) = s(t_i) \times u(t_i) \quad (3.18)$$

The proportion of missing cases calculated in such a way can never be 1, as the survival is correctly estimated since the date of diagnosis. This limitation was overcome by Montanaro *et al.* (2006), who proposed replacing $M(t_i) = 1$ for each time point where $u(t_i) = 1$. This modification allows the method to be applied independently of the actual delay in starting registration, describing a wider set of cancer registry data.

A cancer patient who is dead yet remains unregistered at time T after diagnosis could

have died at any time t_i ($0 < t_i < T$). The proportion of lost cases can thus be derived as:

$$L(t) = \sum_{i=0}^n [s(t_i) - s(t_{i+1})] \cdot [1 - m(t_i)] \cdot u(t_i) \quad (3.19)$$

where $t_n \leq T \leq t_{n+1}$ and t_0 is the time of diagnosis.

Completeness at time T after diagnosis can now be found by subtracting from unity the proportions of missing and lost cases, to get

$$\begin{aligned} C(T) &= 1 - M(T) - L(T) \\ &= 1 - s(t_n) \cdot u(t_n) - \sum_{i=0}^n [s(t_i) - s(t_{i+1})] \cdot [1 - m(t_i)] \cdot u(t_i) \end{aligned} \quad (3.20)$$

In theory, the variance of $C(T)$ could be estimated algebraically, assuming that $s(t_i)$, $m(t_i)$ and $u(t_i)$ are independent. However, bootstrapping offers a more direct approach (Silcocks & Robinson, 2004), providing estimates of the standard error and, roughly, of the form of the sampling distributions for completeness at selected time points. Given the point estimate and the standard error, confidence limits can be found using any appropriate two-parameter distribution. Thus, instead of forcing data into a normal distribution, the observed point estimate can be assumed to follow a beta distribution. Hence, confidence limits can be found from the beta (α, β) distribution with mean equal to the point estimate of completeness θ and variance equal to the square of its standard error σ , derived from the bootstrap. The parameters (α, β) are estimated from the relations:

$$\begin{aligned} (\alpha + \beta) &= \theta(1 - \theta)/\sigma^2 - 1 \\ \alpha &= \theta(\alpha + \beta) \end{aligned} \quad (3.21)$$

The flow method was validated in a simulation modelling study (Silcocks & Robinson, 2007) and it was a major advance in assessing completeness of registration as it describes how completeness increases with time since diagnosis.

Chapter 4

Data and Methods

The goal of this study was to evaluate completeness of registration in RORENO for malignant gastric cancer cases, excluding lymphomas, of patients resident in the district of Porto at date of diagnosis. *In situ* cancers (the ones that involve only the place in which they began and that have not spread) were excluded from the analysis because these tumors are known to be incompletely registered.

In order to apply any of the methods described in Chapter 3, some data must be accessed and treated properly. In this section, the attainment of the databases used and the modifications made to each one are described.

The reason why gastric cancer was chosen to be included in this study is that it is, for both sexes, one of the most common and lethal cancers in Portugal, specially in the North Region (RORENO, 2010), so it is of great interest to evaluate completeness of registration of such tumours.

All the information related to the gastric cancer cases diagnosed in the North Region of Portugal was accessed by the author through RORENO's database, which is kept in a web application, only available to cancer registrars belonging to the health centers in the coverage area. That database comprises the identification of each patient and respective tumour and hospital of origin, as well as information on treatment and follow-ups, relapses, metastases and, if applicable, date and cause of death.

Cause of death is one of the most difficult informations to obtain in the North Region of Portugal, because only the main sources of information (main hospitals) readily provide such information on their patients to RORENO, and it is very hard to get that information from private laboratories. Since some of the methods available for evaluating completeness of registration require that cause of death is known to the registry, the author got in contact with the Northern Health Sub-Region (ARS Norte - Administração Regional de Saúde do Norte), in order to obtain information from death certificates between 2001 and 2004 (data from 2005 and 2006 had already been received by RORENO, as it was mentioned in Section 2.3). By doing so, RORENO's database got five full years of follow-up, knowing for each gastric cancer patient who had died, if a death certificate containing a mention to that cancer had been received.

The death certificates collected were made available in paper, and they were not aggregated by cause of death, so all deaths occurring in the district of Porto had to be verified by the author, one by one, in order to obtain the intended information concerning people who died from gastric cancers or, at least, had a mention to that cancer on the death certificate. The process of gathering data from death certificates, in such a way that those data could be used in this study, took about 2 months.

The data collected from the death certificates contained the identification of the patient (name, date of birth, place of birth and address at the date of death) and the date and cause of death. From 2001 to 2006, 1897 cases were collected from death certificates, which explicitly mentioned gastric cancer and were related to people resident in the district of Porto at the time of death.

After this independent database was built, the information was compared to RORENO's database, in order to complete it. From all death cases collected, 710 did not match directly by name and date of birth, and had to be reinspected. There are many reasons why data may not match: both registration in the web application and collection of death related information were manual processes and, as such, were subject of human error, whether by typing a wrong name or a wrong date of birth; in Portugal, it is very common that women add their husbands' name to their family names when they get married, or

to loose those names when they get divorced, so it is reasonably easy to loose track on those people. Plus, some people who live in the district of Porto at time of death could have lived elsewhere at time of diagnosis and vice-versa. In spite of these difficulties, 308 more cases were successfully traced-back. For 54 other cases, the date of diagnosis could be more accurately defined than just setting it the same as the date of death. However, since no further information could be obtained regarding the hospital where the patient was treated or diagnosed, these cases were treated as DCO cases. The remaining 348 cases were also registered as DCO's.

4.1 Semi-quantitative methods

The qualitative methods used in this study were **M:I ratios** and **histological verification of diagnosis** (MV%).

The first scenario considered for both methods was a comparison between values obtained over time, using cases diagnosed from 2001 to 2006, in order to check for significant differences between those values. In the second scenario, the overall value obtained was compared to some other european registries, for which results for both methods are published in *Cancer Incidence in Five Continents, vol. IX* (Curado *et al.*, 2007). The registries used to make these comparisons were: ROR-Sul, Granada (Spain) and Geneva (Switzerland) – which are regional registries – and Sweeden, Norway, Finland and Denmark – which are national registries.

The analysis was then stratified by sex, for both cenarios, and the same comparisons were made. The results for this and the other methods are presented in Chapter 5.

4.2 Quantitative methods

Regarding quantitative methods, three were applied in this study: capture-recapture, DC and M:I ratio, and the flow method. In all of them, the death certificates collected were used as a source of information, separately from the RORENO's database.

It must be noted that a death certificate may contain information regarding diseases related to a person, and not only the diseases which caused the death. However, sometimes it happens that a cancer is not mentioned in the death certificate, even if it was the cause of

death. To avoid this lack of quality in death certificates, a second scenario was approached in this study, for all quantitative methods, which was to assume that, for every patient known to be dead, a death certificate mentioning cancer was received by the registry. Results for this scenario must be carefully analysed, since this hypothesis is the best case scenario for the evaluation of completeness, leading certainly to an overestimation of that value.

All these applications were also stratified by sex, for both scenarios.

4.2.1 Capture-recapture method

To evaluate completeness of gastric cancer registration using capture-recapture methods, two sources of information were considered: the diagnosed cases from 2001 to 2006, and death certificates collected for the same period.

Among the capture-recapture models described in Chapter 3, and since we are considering two sources of notification, only models M_0 , M_t and M_b can be modelled. In the context of cancer registration, the author chose model M_t as the closest to reality. Thus, a temporal effect was considered in the analysis, causing the capture probabilities to vary among capture occasions. Nevertheless, in order to show how biased the assumption of equality between capture probabilities is, results for model M_0 are also presented.

All calculations for capture-recapture methods were made using package Rcapture (Baillargeon & Rivest, 2009), developed for R environment (R Development Core Team, 2008).

4.2.2 DC and M:I method

Since the DC and M:I method may be seen as a particular case of capture-recapture methods (where the registration of a patient while alive corresponds to the first capture and the registration initiated by the death certificate corresponds to the second capture) the cases diagnosed between 2001 and 2006 were also used in the analysis, as well as the death certificates referent to that period.

Both the expressions available to evaluate completeness through this method were measured, for both scenarios (the real one and the best case scenario, as described above). The results for both expressions are presented in the next chapter, as well as the values

obtained for the parameters

- a – cases registered while alive, not known to have died;
- b – cases registered while alive, known to have died;
- c – cases registered after notifications from death certificates;
- d – cases not registered.

4.2.3 Flow method

The flow method requires that two distinct databases are collected: one containing all cancer cases diagnosed in a given year; and other one containing all deaths registered during a given year (independent from the year of diagnosis).

In this study, the year 2001 was chosen to be used for the diagnosed cases file.

The death certificates collected in a given year help us describe earlier diagnosed cancer patients in further detail, since they provide not only the date and cause of death, but they can also mention for how long a patient had that cancer. Thus, the earlier the year of cancer deaths considered, the more complete information we have on that year, since later years complement that information. Based on this assumption, the death cases file used was related to the year 2001, the first for which death certificates mentioning gastric cancer were collected by the author.

Another fact which supports this decision is that patients with a gastric cancer are known to have a low probability of survival only a year after diagnosis. The closer the years of the diagnosed and the death cases files are, the higher will be the probability of matching data from both files. Hence, since we are using the year 2001 to construct the diagnosed cases file, it is only natural that we use the same year to the death cases file.

Nevertheless, in order to verify that these assumptions holds, a comparative analysis was made using different years for the cancer deaths file, for which results are presented in the next section.

According to the guidelines from the flow method's authors, all data files were checked to verify the absence of:

1. DCOs or DCIs without a date of death;

2. DCOs with date of diagnosis not the same as date of death;
3. Non-DCOs who have date of diagnosis the same as date of death;
4. Cases not belonging to the Registry's region;
5. Non-melanoma skin cancers;
6. Non-malignant tumours;
7. DCOs and DCIs with date of registration preceding date of death;
8. Cases with date of diagnosis later than date of death;
9. Cases aged greater than 120 at diagnosis;
10. Cases with missing values for sex, date of birth, date of diagnosis or date of registration.

However, as it has been said before, only malignant gastric cancer cases, of patients resident in the district of Porto at date of diagnosis, were considered in this study, so the items 4, 5 and 6 did not need further verification.

Table 4.1 shows the number of cases with unknown district of residence and, among them, the ones found later to be from Porto. It also presents the number of cases excluded for violating the guidelines mentioned above, the number of DCO's and the final total number of cases included in each of the files considered in this analysis. The detailed description of these numbers is given below.

	Diagnosed cases file	Death cases file		
	2001	2005	2003	2001
Unknown district of residence	97	15	29	32
Found to be from Porto	13	6	21	15
Cases violating guidelines	18	2	3	3
DCO's	80	68	60	72
Final total	646	488	459	469

Table 4.1: Condensed description of the files used in the flow method application.

Diagnosed cases file

From all cases diagnosed in 2001 on the Registry's database, only the malignant gastric cancers were kept, excluding lymphomas. This file was then cleaned to exclude patients resident in other districts rather than Porto at the date of diagnosis. There were 97 cases with unknown district of residence at diagnosis, 13 of which were found to be resident in Porto through a search in a National Health Database. The file kept 646 cases (80 DCO's), after excluding:

- 1 non-DCO case who had date of diagnosis the same as date of death;
- 1 case aged less than 15 at diagnosis;
- 16 cases with missing values for date of birth.

Death cases files

Death cases from 2005

From all cases with a malignant gastric cancer (except lymphomas) registered in the Registry's database, the ones who died in 2005 were kept. This file was then cleaned to exclude patients resident in other districts rather than Porto at the date of diagnosis. There were 15 cases with unknown district of residence at diagnosis, 6 of which were found to be resident in Porto. The file kept 488 cases (including 68 DCO's), after excluding two non-DCO cases who had date of diagnosis the same as date of death.

Death cases from 2003

From all cases with a malignant gastric cancer (except lymphomas) registered in the Registry's database, the ones who died in 2003 were kept. This file was then cleaned to exclude patients resident in other districts rather than Porto at the date of diagnosis. There were 29 cases with unknown district of residence at diagnosis, 21 of which were found to be resident in Porto. The file kept 459 cases (including 60 DCO's), after excluding three cases with missing values for date of birth.

Death cases from 2001

From all cases with a malignant gastric cancer (except lymphomas) registered in the Registry's database, the ones who died in 2001 were kept. This file was then cleaned to exclude patients resident in other districts rather than Porto at the date of diagnosis. There were 32 cases with unknown district of residence at diagnosis, 15 of which were found to be resident in Porto. The file kept 469 cases (including 72 DCO's), after excluding:

- 1 non-DCO case who had date of diagnosis the same as date of death;
- 1 case aged less than 15 at diagnosis;
- 1 case with missing value for date of birth.

Application

Since two programs have already been written for STATA software by David Robinson (one of the flow method's authors), that software was learned during this study, on version STATA 10 (StataCorp, 2007). The existing programs were then applied to the data mentioned above, after those commands were made available upon the author's request.

It must be emphasized that, when using the data files described above, no DCI information was obtained by further investigation of DCN cases, so only DCO's were kept. In this case, much information is lost, since the flow method does not allow the inclusion of DCO cases in the analysis without the presence of DCI cases.

Through the probability $m(t_i)$, which is the probability that the death certificate of a patient who dies in the time interval (t_i, t_{i+1}) after diagnosis mentions cancer, the flow method also evaluates the quality of information of death certificates. By assuming that, for every patient known to be dead, a death certificate mentioning cancer was received by the registry, the probability $m(t_i)$ will tend to 1 and, by doing so, the proportion of lost cases (L) will tend to 0 (see equation 3.19).

Once again, taking into account that this hypothesis is the best possible scenario for the evaluation of completeness, it was also tested through the flow method, using the year 2001 for both diagnosed and death cases files. The results are presented in the next section.

Chapter 5

Results

In this section, results from the application of different methods available for evaluating completeness of cancer registration are presented. The methods approached were, as mentioned in the previous chapter, the mortality/incidence ratios, histological verification of diagnosis, capture-recapture and both the death certificate methods (DC and M:I ratio and the flow method).

5.1 M:I ratio

The values obtained for mortality:incidence ratios related to the comparisons described in the previous chapter are presented, as well as the number of incident cases and death registered in RORENO's database. For all situations under test, the value obtained for the test statistic is also presented, to draw conclusions about the existence or not of significant differences between results.

5.1.1 Comparative results over time

Tables 5.1, 5.2 and 5.3 show the values of the variables mentioned above, for the comparisons between mortality:incidence ratios obtained over time, from 2001 to 2006, for both sexes, male and female patients, respectively.

While the number of diagnosed cases oscillated around 700 cases each year, the number of deaths registered fluctuated around 500. M:I ratios appear to be decreasing over

Year of diagnosis	Incident cases	Deaths registered	M:I ratio (%)	Z^2
2001	652	497	76.2	0.88
2002	709	512	72.2	0.04
2003	688	518	75.3	0.62
2004	736	541	73.5	0.21
2005	710	476	67.0	0.72
2006	684	433	63.3	2.53

Table 5.1: Comparative results of M:I ratios over time, for both sexes

time, which may indicate an improvement in treating gastric cancer. However, no significant differences were found between the ratios obtained over time, according to the proposed hypotheses in Chapter 3 (3.3), taking into account that Z^2 is always less than 3.84.

Year of diagnosis	Incident cases	Deaths registered	M:I ratio (%)	Z^2
2001	389	306	78.6	1.23
2002	423	309	73.0	0.01
2003	397	307	77.3	0.69
2004	423	320	75.7	0.23
2005	386	263	68.1	1.45
2006	405	277	68.4	1.38

Table 5.2: Comparative results of M:I ratios over time, for male patients

Considering only male patients (Table 5.2), M:I ratios also appear to be decreasing over time, although no significant differences were found between the ratios obtained over time. The lowest value (68.1%) was registered in 2005 and the highest was related to 2001 (78.6%).

Considering only female patients (Table 5.3), the results were similar to the ones obtained for male patients: M:I ratios also appear to be decreasing over time, although no significant differences were found between the ratios obtained over time. The lowest value (55.9%) was registered in 2006 and the highest was related to 2001 (72.6%).

Comparing the results obtained for each sex separately, it can be observed that both the

Year of diagnosis	Incident cases	Deaths registered	M:I ratio (%)	Z^2
2001	263	191	72.6	0.44
2002	286	203	71.0	0.20
2003	291	211	72.5	0.46
2004	313	221	70.6	0.16
2005	324	213	65.7	0.14
2006	279	156	55.9	3.60

Table 5.3: Comparative results of M:I ratios over time, for female patients

numbers of incident cases and of deaths registered were always higher for male than for female patients, and those differences are also reflected in the mortality:incidence ratio. The biggest discrepancy between M:I ratios for each sex is observed in the year 2006, where the M:I ratio obtained for male patients was higher, by 12.5%, than for female patients.

5.1.2 Comparative results with some european registries

For the comparisons between mortality:incidence ratios obtained for different european registries, tables 5.4, 5.5 and 5.6 show the values of the same variables as the ones mentioned above, for both sexes, male and female patients, respectively.

Cancer Registry	Incident cases	Deaths registered	M:I ratio (%)	Z^2
RORENO	4179	2977	71.2	2.52
ROR-Sul	3915	3096	79.1	0.02
Granada (Spain)	585	508	86.8	0.45
Geneva (Switzerland)	207	131	63.3	0.58
Sweeden	5266	4430	84.1	1.89
Norway	3017	2501	82.9	0.67
Finland	3847	2849	74.1	0.84
Denmark	2507	1949	77.7	0.01

Table 5.4: Comparative results of M:I ratios from different registries, for both sexes

Since we are using both regional and national registries, and the countries to which

these registries belong show large differences in the size of the population covered, the number of diagnosed cases and deaths registered also oscillate greatly. Regarding M:I ratios, the highest one was obtained in Granada (86.8%) and the lowest was registered in Geneva (63.3%). Using the test described in Chapter 3 (3.3) for the comparison between M:I ratios obtained for the different registries, no significant differences were found.

Cancer Registry	Incident cases	Deaths registered	M:I ratio (%)	Z^2
RORENO	2423	1781	73.5	1.21
ROR-Sul	2352	1910	81.2	0.59
Granada (Spain)	360	307	85.3	0.45
Geneva (Switzerland)	118	76	64.4	0.57
Sweeden	3182	2619	82.3	1.43
Norway	1825	1486	81.4	0.52
Finland	2106	1542	73.2	1.19
Denmark	1574	1147	72.9	1.03

Table 5.5: Comparative results of M:I ratios from different registries, for male patients

Considering only male patients (Table 5.5), Granada and Geneva registered, once again, the highest (85.3%) and the lowest (64.4%) M:I ratios, respectively. However, no significant differences were found between the ratios obtained for the different registries.

Cancer Registry	Incident cases	Deaths registered	M:I ratio (%)	Z^2
RORENO	1756	1196	68.1	2.95
ROR-Sul	1563	1186	75.9	0.21
Granada (Spain)	225	201	89.3	0.30
Geneva (Switzerland)	89	55	61.8	0.39
Sweeden	2084	1811	86.9	1.64
Norway	1192	1016	85.2	0.59
Finland	1741	1307	75.1	0.37
Denmark	933	801	85.9	0.55

Table 5.6: Comparative results of M:I ratios from different registries, for female patients

Considering only female patients (Table 5.6), the results were similar to the ones obtained for male patients and for both sexes: the lowest value (61.8%) was registered in

Geneva and the highest was related to Granada (89.3%). No significant differences between M:I ratios were found in this analysis.

Comparing the results obtained for each sex separately, it can be observed that both the numbers of incident cases and of deaths registered were always higher for male than for female patients, but those differences were not reflected in the mortality:incidence ratios, e.g., Finland registered a M:I ratio of 73.2% for male patients, which was lower than the ratio of 75.1% obtained for female patients. The biggest discrepancy between M:I ratios for each sex is observed in Denmark, where the M:I ratio obtained for male patients was lower, by 13%, than for female patients.

It must be noticed that the M:I ratios published in *Cancer Incidence in Five Continents vol. IX* (Shin *et al.*, 2007) are related to the period of 1998-2002 (Table 2.1), instead of the period 2001-2006, considered in the M:I ratio values obtained for RORENO in this study. However, since no significant differences over time were found between RORENO's M:I ratios, and since one time period is partially covered by the other, the author considered that those values could be compared.

5.2 Histological verification of diagnosis

Similarly to the analysis of the previous method, the values obtained for the proportion of cases with histological verification of diagnosis (i.e., with diagnosis being made through histology or cytology/haematology analysis), related to the comparisons described in the previous chapter are presented, as well as the number of incident cases and the number of cases microscopically verified in RORENO's database. For all situations under test, the test statistic is also presented, to draw conclusions about the existence or not of significant differences between results.

5.2.1 Comparative results over time

Tables 5.7, 5.8 and 5.9 show the values of the variables mentioned above, for the comparisons between the proportion of microscopically verified cases obtained over time, from 2001 to 2006, for both sexes, male and female patients, respectively.

Year of diagnosis	Incident cases	Microscopically verified	$MV\%$	Z^2
2001	652	549	84.2	0.66
2002	709	579	81.7	2.03
2003	688	620	90.1	0.24
2004	736	647	87.9	0.00
2005	710	635	89.4	0.12
2006	684	644	94.2	1.95

Table 5.7: Comparative results of $MV\%$ over time, for both sexes

As we have seen before, the number of diagnosed cases oscillated around 700 cases each year, while the number of microscopically verified cases fluctuated around 600. The proportion of MV cases does not appear to have a regular evolution over time. Moreover, no significant differences were found between the proportions obtained over time, according to the proposed hypotheses in Chapter 3 (3.10).

Considering only male patients (Table 5.8), $MV\%$ appear to be increasing over time, although no significant differences were found between the proportions obtained over time. The lowest value (81.3%) was registered in 2002 and the highest was related to 2006 (95.1%).

Year of diagnosis	Incident cases	Microscopically verified	<i>MV%</i>	Z^2
2001	389	333	85.6	0.36
2002	423	344	81.3	2.35
2003	397	352	88.7	0.00
2004	423	377	89.1	0.01
2005	386	355	92.0	0.47
2006	405	385	95.1	1.81

Table 5.8: Comparative results of *MV%* over time, for male patients

Considering only female patients (Table 5.9), the results were similar to the ones obtained for male patients: *MV%* also appear to be increasing over time – with the exception of year 2003, which presents a high value for *MV%* – although no significant differences were found between the ratios obtained over time. The lowest value (82.1%) was registered in 2001 and the highest was related to 2006 (92.8%).

Year of diagnosis	Incident cases	Microscopically verified	<i>MV%</i>	Z^2
2001	263	216	82.1	1.04
2002	286	235	82.2	1.11
2003	291	268	92.1	1.24
2004	313	270	86.3	0.03
2005	324	280	86.4	0.02
2006	279	259	92.8	1.56

Table 5.9: Comparative results of *MV%* over time, for female patients

Comparing the results obtained for each sex separately, it can be observed that both the numbers of incident cases and of microscopically verified cases were always higher for male than for female patients, but those differences did not reflect in the proportion *MV%*. The biggest discrepancy between *MV%* for each sex is observed in the year 2005, where the *MV%* obtained for male patients was 5.6% higher than for female patients.

5.2.2 Comparative results with some european registries

For the comparisons between the proportion of microscopically verified cases obtained for different european registries, tables 5.10, 5.11 and 5.12 show the values of the same variables as the ones mentioned above, for both sexes, male and female patients, respectively.

Cancer Registry	Incident cases	Microscopically verified	MV%	Z^2
RORENO	4179	3675	87.9	0.67
ROR-Sul	3915	3212	82.0	3.46
Granada (Spain)	585	527	90.1	0.02
Geneva (Switzerland)	207	204	98.6	0.07
Sweeden	5266	5225	99.2	2.11
Norway	3017	2883	95.6	0.27
Finland	3847	3686	95.8	0.40
Denmark	2507	2303	91.9	0.00

Table 5.10: Comparative results of MV% from different registries, for both sexes

As in the M:I ratio method, the number of diagnosed cases and the number of microscopically verified cases oscillate greatly. Regarding MV%, the highest one was obtained in Sweeden (99.2%) and the lowest was registered in ROR-Sul (82.0%). Using the test described in Chapter 3 (3.10) for the comparison between MV proportions obtained for the different registries, no significant differences were found.

Considering only male patients (Table 5.11), Sweeden and ROR-Sul registered, once again, the highest (99.5%) and the lowest (83.8%) MV proportions, respectively. However, no significant differences were found between the ratios obtained for the different registries.

Considering only female patients (Table 5.12), the lowest value (79.4%) was also registered in ROR-Sul but the highest was related to Geneva (98.9%), although the difference between Geneva and Sweeden is quite small (only 0.1%). Once again, no significant differences were found between the proportions MV% obtained from the different registries.

Cancer Registry	Incident cases	Microscopically verified	<i>MV%</i>	Z^2
RORENO	2423	2147	88.6	0.81
ROR-Sul	2352	1971	83.8	3.32
Granada (Spain)	360	332	92.2	0.00
Geneva (Switzerland)	118	116	98.3	0.05
Sweeden	3182	3166	99.5	2.09
Norwaya	1825	1761	96.5	0.33
Finland	2106	2032	96.5	0.38
Denmark	1574	1458	92.6	0.01

Table 5.11: Comparative results of *MV%* from different registries, for male patients

Cancer Registry	Incident cases	Microscopically verified	<i>MV%</i>	Z^2
RORENO	1756	1528	87.0	0.50
ROR-Sul	1563	1241	79.4	3.63
Granada (Spain)	225	195	86.7	0.08
Geneva (Switzerland)	89	88	98.9	0.09
Sweeden	2084	2059	98.8	2.08
Norway	1192	1122	94.1	0.18
Finland	1741	1654	95.0	0.44
Denmark	933	845	90.6	0.00

Table 5.12: Comparative results of *MV%* from different registries, for female patients

Comparing the results obtained for each sex separately, it can be observed that both the numbers of incident cases and of microscopically verified cases were, once more, always higher for male than for female patients. The only registry for which those differences were not reflected in the proportion *MV%* was Geneva, yielding a *MV%* for male patients lower, by 0.6%, than for female patients. The biggest discrepancy between *MV%* for each sex was observed in Granada, where the *MV* proportion obtained for male patients was 5.5% higher than for female patients.

5.3 Capture-recapture method

In the application of capture-recapture methods, two models were applied, M_0 and M_t , as it was mentioned before. The tables presented in this section show the values obtained for the capture probabilities estimated for each model, the estimates for the population size, the AIC – Akaike Information Criterion, proposed by Akaike (1974) – and the value for completeness, with the respective confidence intervals using 95% confidence level. Both the scenarios described in the previous chapter were applied, and the analysis was stratified by sex.

5.3.1 First scenario

This scenario takes into consideration the real data, of cases registered in RORENO's database between 2001 and 2006, and of the death certificates collected regarding the same period. The values of the variables mentioned above are presented in Tables 5.13 and 5.14.

Model M_0					
Sex	$\hat{N}$	Estimated parameters	AIC	Completeness	95% CI
Male	3823	p=0.39	1286.7	63.4	[60.3 ; 66.5]
Female	2340	p=0.50	615.2	75.1	[72.0 ; 78.1]
Both sexes	6078	p=0.44	1860.1	68.8	[66.6 ; 70.9]

Table 5.13: Completeness values and respective confidence intervals obtained by adjusting model M_0 , in the first scenario.

The overall completeness of gastric cancer registration, using model M_0 , was evaluated in 68.8%, and was found to be lower for male patients (63.4%) than for female patients (75.1%). Only one value of capture probability was estimated, since model M_0 assumes that the capture probability is constant over time and for all individuals, as it has been mentioned in Chapter 3.

Using model M_t , overall completeness was evaluated in 82.9%, and was also found to be lower for male patients (81.5%) than for female patients (84.5%). The capture probabilities estimated indicate that it is more likely to capture an individual by registration

Model M_t					
Sex	$\hat{N}$	Estimated parameters	AIC	Completeness	95% CI
Male	2976	p1=0.75 ; p2=0.27	30.6	81.5	[78.6 ; 84.2]
Female	2072	p1=0.76 ; p2=0.37	30.0	84.5	[82.0 ; 87.4]
Both sexes	5042	p1=0.75 ; p2=0.31	32.4	82.9	[80.9 ; 84.8]

Table 5.14: Completeness values and respective confidence intervals obtained by adjusting model M_t , in the first scenario.

while alive (first capture occasion) than by a death certificate (second capture occasion), since $p_1 > p_2$ in every case.

One must notice the great differences between the values obtained for AIC for each model, which shows that model M_t (with lower values for AIC) illustrates the reality better than model M_0 , as it was supposed.

5.3.2 Second scenario

This scenario may be considered as the best case scenario, as it uses the cases registered in RORENO's database between 2001 and 2006 and, for all of them known to have died, it assumes that a death certificate is received by the registry. The values of the variables mentioned in the beginning of this section are presented in Tables 5.15 and 5.16.

Model M_0					
Sex	$\hat{N}$	Estimated parameters	AIC	Completeness	95% CI
Male	2535	p=0.79	267.5	95.6	[94.5 ; 96.5]
Female	1895	p=0.73	220.9	92.7	[90.1 ; 94.2]
Both sexes	4424	p=0.76	461.9	94.5	[93.6 ; 95.3]

Table 5.15: Completeness values and respective confidence intervals obtained by adjusting model M_0 , in the second scenario.

The overall completeness of gastric cancer registration, using model M_0 , was evaluated in 94.5%, and was found to be lower for female patients (92.7%) than for male patients (95.6%). Again, only one value of capture probability was estimated, since model M_0 assumes that the capture probability is constant over time and for all individuals.

Model M_t					
Sex	$\hat{N}$	Estimated parameters	AIC	Completeness	95% CI
Male	2505	p1=0.89 ; p2=0.71	30.7	96.7	[95.8; 97.6]
Female	1861	p1=0.74 ; p2=0.64	30.0	94.4	[92.9 ; 96.7]
Both sexes	4361	p1=0.87 ; p2=0.68	32.4	95.8	[95.1 ; 96.6]

Table 5.16: Completeness values and respective confidence intervals obtained by adjusting model M_t , in the second scenario.

Using model M_t , overall completeness was evaluated in 95.8%, and was also found to be lower for female patients (94.4%) than for male patients (96.7%). The capture probabilities estimated still indicate that it is more likely to capture an individual by registration while alive (first capture occasion) than by a death certificate (second capture occasion), since $p1 > p2$ in every case, although the difference between p1 and p2 is less prominent, as it was expected.

Once again, one must notice the great differences between the values obtained for AIC for each model, which shows that model M_t (with lower values for AIC) illustrates the reality better than model M_0 .

5.4 DC and M:I method

For the DC and M:I method, two tables are presented, containing the values of the parameters a , b , c and d obtained, the proportion of DCI cases and the values of completeness, evaluated using both formulas (3.16) and (3.17) given in Chapter 3.

5.4.1 First scenario

Using Ajiki's formula (3.17), $\frac{1-DCI \times (1/M:I)}{1-DCI}$, overall completeness was evaluated in 82.9%, and was found to be higher for female patients (84.7%) than for male patients (81.4%).

Sex	a	b	c	$\hat{d}$	DCI (%)	Ajiki Comp. (%)	Parkin Comp. (%)
Male	596	1624	203	553	8.4	81.4	85.4
Female	586	981	189	316	10.8	84.7	88.0
Both sexes	1182	2605	392	864	9.4	82.9	86.6

Table 5.17: DCI cases and completeness values obtained in the first scenario.

Since the proportion of DCI cases for female patients exceeds 10%, the evaluation of completeness by sex using Parkin's formula (3.16), $\frac{1}{1-DCI+DCI \times (1/M:I)}$, is compromised. Nevertheless, for both sexes, it could be assumed that the overall completeness was valued 86.6%.

5.4.2 Second scenario

Sex	a	b	c	$\hat{d}$	DCI (%)	Ajiki Comp. (%)	Parkin Comp. (%)
Male	1579	641	203	82	8.4	96.7	97.1
Female	1006	561	189	105	10.8	94.3	95.2
Both sexes	2585	1202	392	182	9.4	95.8	96.4

Table 5.18: DCI cases and completeness values obtained in the second scenario.

As it was expected in the second scenario (Table 5.18), the number of registered cases while alive and with information on the death certificate (a) greatly expanded, opposedly to the number of registered cases while alive and with no information on the death certificate (b). This obviously led to a big drop on the number of missed cases (d), thus leading to the expected overestimation of the overall completeness, which was evaluated in 95.8%, using Ajiki's formula, and 96.4%, using Parkin's formula.

Since the proportion of DCI cases does not change using the second scenario, the use of Parkin's formula for estimating completeness by sex is still of limited use, since DCI% for female patients exceeds 10%.

Similarly to what had been verified in the capture-recapture method, completeness for each sex separately became higher for male patients than for female patients.

5.5 Flow method

As it has been mentioned before, two scenarios were taken into account in order to apply the flow method. In the first scenario, three different data sets were used for the death cases file. The second scenario, which may be considered as the best case scenario, it is assumed that, for every cancer patient that died, a death certificate mentioning cancer was received. The diagnosed cases file used was related to the year 2001, for both scenarios. In order to get a broader understanding of these results, completeness of registration was also evaluated for each sex separately, for each scenario. Results are presented below.

5.5.1 First scenario

Figures 5.1, 5.2 and 5.3 show the completeness curves, as well as the proportion of lost cases through time, respectively considering the years 2005, 2003 and 2001 for the death cases file. The obtained values for completeness of registration at 1, 2, 3, 4 and 5 years after diagnosis and the corresponding 95% confidence intervals are presented in tables 5.19, 5.21 and 5.23 once again considering the years 2005, 2003 and 2001 for the death cases file, respectively.

Completeness of registration for both men and women was evaluated at 1, 2, 3, 4 and 5 years after diagnosis and the results are presented on tables 5.20, 5.22 and 5.24.

Death cases from 2005

Time after diagnosis	Completeness	95% CI
1 year	34.34	[20.56 ; 49.62]
2 years	41.36	[24.80 ; 59.00]
3 years	44.32	[26.54 ; 62.89]
4 years	47.67	[27.71 ; 65.20]
5 years	46.17	[27.93 ; 66.31]

Table 5.19: Completeness of registration using deaths registered in 2005

Through the analysis of Table 5.19 and Figure 5.1, one may notice the evolution of completeness over time, using the year 2005 for the death cases file. As it was expected,

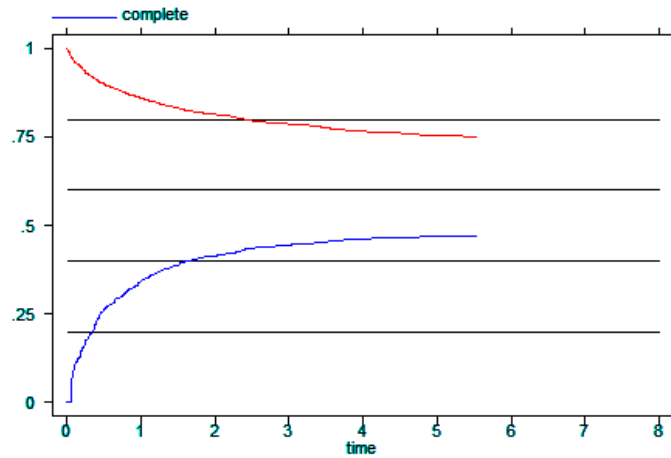


Figure 5.1: Completeness plus proportion lost, using deaths registered in 2005

the value obtained for completeness is extremely low, and the width of the 95% confidence intervals is very high, reaching almost 40% at five years after diagnosis.

Time after diagnosis	Men		Women	
	Completeness	95% CI	Completeness	95% CI
1 year	34.07	[10.36 ; 63.43]	35.03	[8.61 ; 68.35]
2 years	40.88	[11.94 ; 74.12]	42.40	[9.64 ; 79.95]
3 years	43.57	[12.53 ; 77.97]	45.63	[10.14 ; 84.21]
4 years	45.64	[12.84 ; 80.93]	46.41	[10.05 ; 85.43]
5 years	46.28	[12.13 ; 82.75]	47.07	[8.90 ; 87.69]

Table 5.20: Completeness of registration according to sex, using deaths registered in 2005

The values obtained for each sex, separately, show even wider confidence intervals (Table 5.20), leading to distrust these results. However, it must be noted that completeness of registration for female patients was always higher than for men, being in agreement with the other quantitative methods.

Death cases from 2003

Time after diagnoses	Completeness	95% CI
1 year	44.35	[37.07 ; 51.76]
2 years	51.58	[43.97 ; 59.15]
3 years	53.94	[46.65 ; 61.15]
4 years	55.42	[48.23 ; 62.49]
5 years	55.98	[48.38 ; 63.44]

Table 5.21: Completeness of registration using deaths registered in 2003

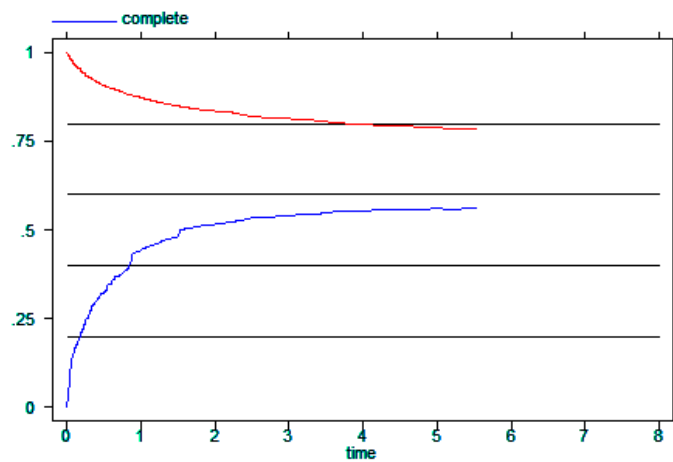


Figure 5.2: Completeness plus proportion lost, using deaths registered in 2003

As it was expected, using year 2003 for the dead cases file led to higher values of completeness than using year 2005 and the confidence intervals also became narrower. Completeness at five years after diagnosis was estimated in 55.98% (Table 5.21) and the proportion of lost cases decreased (Figure 5.2).

The stratified analysis by sex (Table 5.22) yielded the same conclusions as before: female patients presented higher values of completeness than male patients. Confidence intervals were also narrower than using year 2005 for the death cases file.

Time after diagnosis	Men		Women	
	Completeness	95% CI	Completeness	95% CI
1 year	45.96	[34.19 ; 57.96]	42.02	[19.73 ; 66.22]
2 years	51.37	[41.02 ; 61.67]	52.15	[23.21 ; 80.29]
3 years	53.51	[43.69 ; 63.20]	54.74	[24.67 ; 82.96]
4 years	55.16	[45.85 ; 64.29]	55.37	[24.67 ; 83.88]
5 years	55.66	[46.40 ; 64.72]	55.90	[23.26 ; 85.79]

Table 5.22: Completeness of registration according to sex, using deaths registered in 2003

Death cases from 2001

Time after diagnoses	Completeness	95% CI
1 year	76.08	[65.60 ; 85.18]
2 years	79.58	[70.07 ; 87.67]
3 years	80.37	[71.29 ; 88.07]
4 years	83.00	[73.66 ; 90.63]
5 years	83.14	[74.01 ; 90.62]

Table 5.23: Completeness of registration using deaths registered in 2001

When using 2001 as the year of the dead cases file, the values obtained for completeness increased drastically (Table 5.23), rising up to the value which had also been found through the application of the other quantitative methods, 83%, at five years after diagnosis. It must also be noted that, only one year after diagnosis, completeness of registration reached 76%.

Once again, the stratified analysis by sex (Table 5.24) yielded that female patients presented higher values of completeness than male patients. These results vindicate the assumptions made in Section 4.2.3 and motivate the use of year 2001 in the dead cases file, since it provides the most reasonable and credible results of completeness.

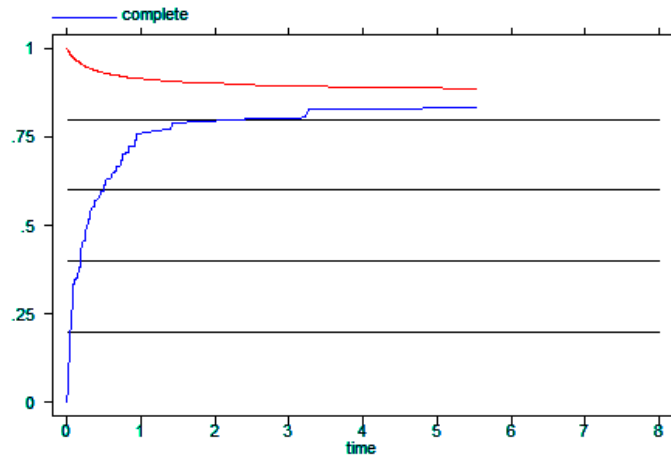


Figure 5.3: Completeness plus proportion lost, using deaths registered in 2001

Time after diagnosis	Men		Women	
	Completeness	95% CI	Completeness	95% CI
1 year	71.48	[57.74 ; 83.45]	83.60	[58.15 ; 98.05]
2 years	76.36	[63.01 ; 87.48]	85.11	[63.16 ; 97.85]
3 years	77.17	[64.48 ; 87.74]	85.82	[67.76 ; 97.07]
4 years	77.80	[65.71 ; 87.87]	89.53	[71.70 ; 98.87]
5 years	77.98	[66.13 ; 87.88]	89.60	[73.24 ; 98.59]

Table 5.24: Completeness of registration according to sex, using deaths registered in 2001

5.5.2 Second scenario

Time after diagnoses	Completeness	95% CI
1 year	84.63	[73.90 ; 92.90]
2 years	89.46	[80.56 ; 95.85]
3 years	90.97	[83.23 ; 96.49]
4 years	94.01	[86.63 ; 98.53]
5 years	94.39	[87.39 ; 98.65]

Table 5.25: Completeness of registration assuming that death certificates are received for every patient that died

Similarly to the other quantitative methods, overall completeness at five year after

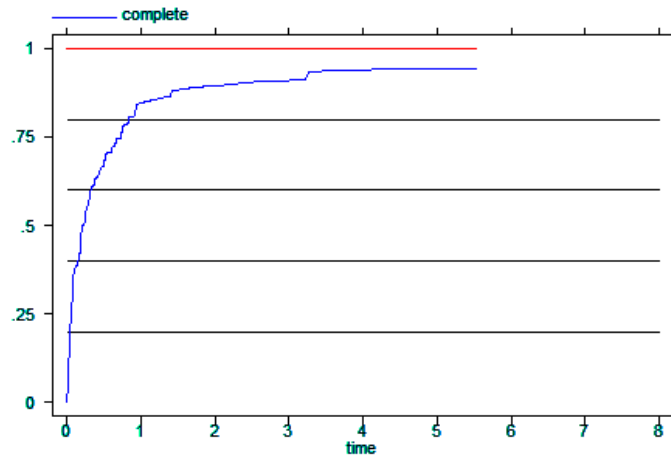


Figure 5.4: Completeness plus proportion lost, assuming that death certificates are received for every patient that died

diagnosis was evaluated in 94.4% (Table 5.25). As it has been explained before, the proportion of lost cases was considered to be zero, which can be observed in Figure 5.4.

Time after diagnosis	Men		Women	
	Completeness	95% CI	Completeness	95% CI
1 year	82.28	[69.92 ; 91.90]	88.73	[67.15 ; 99.21]
2 years	88.75	[78.89 ; 95.78]	91.22	[74.09 ; 99.37]
3 years	90.27	[81.31 ; 96.52]	92.68	[77.53 ; 99.56]
4 years	91.57	[83.43 ; 97.12]	96.62	[81.19 ; 100.00]
5 years	92.08	[84.53 ; 97.25]	96.85	[82.61 ; 100.00]

Table 5.26: Completeness of registration according to sex, assuming that death certificates are received for every patient that died

In this case, contrarily to what had been noticed in the other quantitative methods, overall completeness for female patients (96.9%) remained higher than for male patients (92.1%), being in agreement to the results obtained through the first scenario.

Chapter 6

Conclusions

Cancer registration is a powerful tool for monitoring trends and promoting the necessary clinical and epidemiological research (Schouten, 1996).

The future of cancer registration lies on the continued collaboration between clinicians and epidemiological researchers and in collaborative clinical epidemiological study designs.

The goal of this study was to evaluate completeness of registration in the North Regional Cancer Registry of Portugal for malignant gastric cancer cases, excluding lymphomas, of patients resident in the district of Porto at date of diagnosis. In order to accomplish this task, two databses were collected: one containing the gastric cancer cases registered in RORENO between 2001 and 2006; and another containing data collected from death certificates, regarding the same period.

It is known that, sometimes, data from a cancer registry is incomplete or inaccurate, but it also happens that a death certificate is included in the analysis as a DCO case when, if fact, it did not belong to the coverage area of the registry at time of diagnosis, or the cause of death was misclassified in the death certicate. All of these situations brought difficulties to the collection of data from death certificates to use in this study, further complicated by the fact that the information on death certificates is not computerized.

No significant differences were found between the values obtained by the semi-quantitative methods applied in this study (M:I ratios and MV%), either over time or by comparison

with other European registries, considering both sexes or each sex separately. Hence, one may assume that completeness of gastric cancer registration in RORENO has been consistent with the values obtained by the european registries considered in the analysis (some of which are the oldest in Europe), and that it did not change considerably over time. Nevertheless, one must notice that there has been a decrease in the values of M:I ratios obtained over time, which may indicate an improvement in health care related to cancer treatment.

Regarding quantitative methods, the overall estimates for completeness of gastric registration, using the different methods, gave approximately the same value of 83%, rising to a value of approximately 95%, when using the best case scenario. The results also showed that, when using the real data scenario, female patients presented higher values for completeness of registration (around 85%) than male patients (around 81%), for all methods.

It must be noted that quantitative methods are based on some assumptions that are, generally, not true in cancer registration, and applying these methods usually lead to an undestimation of the real value for completeness:

- In the capture-recapture application: dependence between sources of information was modelled using log-linear models, but heterogeneity between individuals was not taken into account in the estimation of capture probabilities, which may also lead to underestimation of the overall completeness values.
- In the DC and M:I method: the fact that not every death certificate regarding a cancer patient is received by the registry biases the parameters a and b , leading to an overestimation of the number of cases missed by the registry, which causes completeness to be underestimated.
- In the flow method application: it is thought that survival of gastric cancer in the North Region of Portugal is somewhat overestimated. Supposing that survival was in fact lower than the one estimated in this method, would decrease both the proportion of missed (M) and lost (L) cases (equations 3.18 and 3.19, respectively), which would increase the value for completeness.

When using the second scenario in the capture-recapture method and in the DC and M:I ratio method, completeness of registration was evaluated as being higher for male patients than for female patients, which did not occur in the first scenario. An explanation to this fact is that these methods take into consideration the absolute number of individuals in a specific situation, in order to construct a table as in Figure 3.1 or Figure 3.3. In this study specifically, the number of male patients, known to have died, but for which no information was obtained from death certificates (682 individuals), greatly exceeded the number of female patients in the same situation (420 individuals). So, when the assumption related to the second scenario was made, overall results of completeness for each sex were reversed.

The flow method was found to be quite sensitive, in this study, to the choice of the year used for the death cases file. This is due to the fact that death certificates were only collected until 2006, so only 2001 had full follow-up information for five years (whereas 2002 had four years of complete follow-up, 2003 had three years, and so on). Nevertheless, flow method still presented some advantages over the other quantitative methods: when applying the best case scenario to the data, the overall completeness for female patients remained higher than for male patients; and it provided an evaluation of completeness over time, which did not happen with any other method.

Easier access to individual cause of death should be granted to cancer registrars and medical practitioners at large, in order to allow for the use of the methods described in this study and also to improve the quality of data collected, concerning not only completeness of registration, but also accuracy of information, which is another major quality indicator.

A previous study has been published (Pinheiro *et al.*, 2002) containing results for completeness of registration in Portugal using semi-quantitative methods. However, the results presented in this study, obtained through quantitative methods, are the first ones to be published in Portugal.

Although only gastric cancer was considered here, there is an imperious need to apply these methods more often to assess the quality of data in the cancer registries of Portugal. Another issue that would worth being developed in a future work is a methodology to validate the DC and M:I ratio method, since no quality measure related to this method was found in literature.

Appendix

M:I ratio

After defining the vectors of size n , d and m , of the number of diagnosed cases and of deaths registered in each of the n registries, respectively, we set:

```
> n=length(d)
> num=d+m
> num
> teta=sum(m)/sum(d)
> phi=1/(n-1)*sum((m-teta*d)^2/(num*teta))
```

A vector containing the test statistic, for each of the registries under test, is given by:

```
> z=NULL
> for (j in 1:n){
>   z2=(m[j]-teta*d[j])^2/(phi*num[j]*teta)
>   z=c(z,round(z2,2))
> }
> z
```

A registry is marked as unusual if its test statistic is equal to or greater than 3.84, as it has been mentioned in Section 3.1.2.

MV%

After defining the vectors of size n , d and y , of the number of diagnosed cases and of microscopically verified cases in each of the n registries, respectively, we set:

```

> p=sum(y)/sum(d)
> p
> phi=1/(n-1)*sum((y-p*d)^2/(d*p*(1-p)))
> phi

```

A vector containing the test statistic, for each of the registries under test is given by:

```

> z=NULL
> for (j in 1:n){
>   z2=(y[j]-p*d[j])^2/(phi*p*(1-p)*d[j])
>   z=c(z,round(z2,2))
> }
> print(z)

```

A registry is marked as unusual if its test statistic is equal to or greater than 3.84, as it has been mentioned in Section 3.1.4.

Capture-recapture method

We start by loading package Rcapture from R library:

```

> library(Rcapture)

```

After defining wanted n_{11} , n_{10} and n_{01} , as in Figure 3.1 we set:

```

> total=n11+n01+n10
> x=matrix(ncol=3,nrow=3, c(1, 1, n11, 1, 0, n10, 0, 1, n01), byrow=T)
> colnames(x)=c("registro", "dc", "freq")

```

The overall results by capture-recapture methods may be obtained through:

```

> calc=closedp.t(x, dfreq=T)
> print(calc)

```

from which parameters estimates are found using:

```

> calc$parameters

```

In order to obtain overall completeness and 95% confidence intervals for models M_0 and M_t we defined:

```
> IC0=closedpCI.t(x, dfreq=T, m="M0")
> ICt=closedpCI.t(x, dfreq=T, m="Mt")

> round(c(total/IC0$CI[, "abundance"], total/IC0$CI[, "SupCL"],
total/IC0$CI[, "InfCL"]), 4)
> round(c(total/ICt$CI[, "abundance"], total/ICt$CI[, "SupCL"],
total/ICt$CI[, "InfCL"]), 4)
```

The two last commands provide a vector containing overall completeness, the lower limit and the upper limit of the 95% confidence intervals, for models M_0 and M_t , respectively.

DC and M:I method

After setting a , b and c according to Figure 3.3, one estimates d through

```
> d=b*c/a
```

and completeness can simply be estimated using equation (3.15)

```
> comp=(a+b+c) / (a+b+c+d)
> comp
```

or, using Ajiki's formula (3.17),

```
> DCI=c / (a+b+c)
> MI=(a+c) / (a+b+c)
> comp=(1-DCI/MI) / (1-DCI)
> comp
```

If $DCI < 10\%$, one may use Parkin's formula (3.16):

```
> DCI
> comp2=1 / ( (1-DCI) + (DCI/MI) )
> comp2
```

References

- Ajiki, W., Tsukuma, H., & Oshima, A. 1998. Index for evaluating completeness of registration in population-based cancer registries and estimation of registration rate at the Osaka Cancer Registry between 1966 and 1992 using this index. *Nippon Koshu Eisei Zasshi*, **45**, 1011–1017.
- Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.
- Association, World Medical. 2008. *WMA Declaration of Helsinki*. World Medical Association.
- Baillargeon, Sophie, & Rivest, Louis-Paul. 2009. *Rcapture: Loglinear Models for Capture-Recapture Experiments*. R package version 1.2-0.
- Bishop, Y., Fienberg, S., & Holland, P. 1975. *Discrete multivariate analysis: theory and practice*.
- Bray, F., & Parkin, D. M. 2009. Evaluation of data quality in the cancer registry: Principles and methods Part I: Comparability, validity and timeliness. *European Journal of Cancer*, **45**(5).
- Brenner, H. 1995. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*, **6**(1), 42–8.
- Brenner, H., Stegmaier, C., & Ziegler, H. 1994. Estimating completeness of cancer registration in Saarland/Germany with capture-recapture methods. *European Journal of Cancer*, **30**(11), 1659 – 1663.

- Brenner, H., Stegmaier, C., & Ziegler, H. 1995. Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany. *J Epidemiol Community Health*, **49**(4), 426–30.
- Brewster, D. H., & Stockton, D. L. 2008. Ascertainment of breast cancer by the Scottish Cancer Registry: An assessment based on comparison with five independent breast cancer trials databases. *The Breast*, **17**(1), 104 – 106.
- Bullard, J., Coleman, M. P., Robinson, D., Lutz, J. M., Bell, J., & Peto, J. 2000. Completeness of cancer registration: a new method for routine use. *Br J Cancer*, **82**(5), 1111–1116.
- Clemmesen, J. 1974. Statistical studies in the aetiology of malignant neoplasms. *Acta Pathol Microbiol Scand Suppl.*, **Suppl 247**, 1–266.
- Curado, M. P., Edwards, B., Shin, H. R., Storm, H., Ferlay, J., Heanue, M., & Boyle, P. 2007. *Cancer incidence in five continents*. Vol. IX (IARC Scientific Publication No. 160). Lyon: IARC (WHO) and IACR.
- Doll, R., Payne, P., & Waterhouse, J. 1966. *Cancer incidence in five continents*. Vol. I. Springer-Verlag.
- ENCR. 2003. *Standards and guidelines for cancer registration in Europe*. (IARC Technical Publication No. 40). Lyon: IARC(WHO) and ENCR.
- Esteban, D., Whelan, S., Laudico, A., & Parkin, D.M. 1995. *Manual for Cancer Registry Personnel*. (IARC Technical Report No. 10). Lyon: IARC (WHO) and IACR.
- Goldberg, J., Gelfand, H.M., & Levy, P.S. 1980. Registry Evaluation Methods: A Review and Case Study. *Epidemiological Reviews*, **2**, 210–220.
- Hay, G. 1997. The selection from multiple data sources in epidemiological capture-recapture studies. *The Statistician*, **46**(4).
- IARC. 2004. *Guidelines on confidentiality for population-based cancer registration*. (Internal report No. 2004/03). Lyon: IARC(WHO) and IACR.

- Ingelfinger, J.R., & Drazen, J.M. 2004. Registry research and medical privacy. *N Engl J Med*, **350**, 1452–1453.
- Jensen, O.M., Storm, H.H., & Jensen, H.S. 1985. Cancer registration in Denmark and the study of multiple primary cancers, 1943–80. *Natl Cancer Inst Monogr*, **68**, 245–251.
- Kamo, K., Kaneko, S., Satoh, K., Yanagihara, H., Mizuno, S., & Sobue, T. 2007. A mathematical estimation of true cancer incidence using data from population-based cancer registries. *Japanese Journal of Clinical Oncology*, **37**(2), 150–155.
- Larsen, I. K., Smastuen, M., Johannesen, T. B., Langmark, F., Parkin, D. M., Bray, F., & Moller, B. 2009. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. *Eur J Cancer*, **45**(7), 1218–31.
- Montanaro, F., Robinson, D., Bordoni, A., & Lutz, J-M. 2006. A modification to the flow method to estimate completeness in cancer registries with delayed registration. *J Public Health*, **28**(3), 274–7.
- Morabia, A. 2004. *History of epidemiological methods and concepts*. Birkhäuser.
- Parkin, D. M. 2006. The evolution of the population-based cancer registry. *Nature Reviews Cancer*, **6**(8), 603–612.
- Parkin, D. M., & Bray, F. 2009. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *Eur J Cancer*, **45**(5), 756 – 764.
- Parkin, D. M., & Plummer, M. 2002. Comparability and quality of data. *Chap. 5, pages 57–73 of: Parkin, D. M., Whelan, S. L., Ferlay, J., Teppo, L., & Thomas, D. B. (eds), Cancer incidence in five continents., vol. VIII (IARC Scientific Publication No. 155). Lyon: IARC.*
- 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: IARC (WHO) and IACR.
- Parkin, D. M., Wabinga, H., & Nambooz, S. 2001. Completeness in an African cancer registry. *Cancer Causes Control*, **12**(2), 147–52.

- Pinheiro, P.S., Tyczynski, J.E., Bray, F., Amado, J., Matos, E., Miranda, A.C., & Limbert, E. 2002. *Cancer in .. Cancro em Portugal (IARC Technical Publication No. 38)*. Lyon: IARC.
- R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Regal, R.R., & Hook, E.B. 1991. The effects of model selection on confidence intervals for the size of a closed population. *Stat Med*, **10**, 717–721.
- Rivest, L.P., & Lévesque, T. 2001. Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics*, **29**(4), 555–572.
- ROR-Sul. 2008. *A história do registo de cancro em Portugal: Um contributo pessoal*. Registo Oncológico Regional do Sul, ed.
- RORENO. 2010. *Análise de Sobrevivência: Principais Cancros da Região Norte 2000/2001*. Registo Oncológico Regional do Norte, ed.
- Ruiz, A., & Facio, A. 2004. Hospital-based cancer registry: A tool for patient care, management and quality. A focus on its use for quality assessment. *Clinical and Translational Oncology*, **6**(2), 104–113.
- Schmidtman, I. 2008. Estimating completeness in cancer registries—comparing capture-recapture methods in a simulation study. *Biom J*, **50**(6), 1077–92.
- Schouten, L.J. 1996. *Cancer Registration: data quality and prospects for use*. Heythuy-sen.
- Shin, H. R., Curado, M. P., Ferlay, J., Heanue, M., Edwards, B., & Storm, H. 2007. Comparability and quality of data. *Chap. 5, pages 67–94 of: Curado, M. P., Edwards, B., Shin, H. R., Storm, H., Ferlay, J., Heanue, M., & Boyle, P. (eds), Cancer incidence in five continents., vol. IX (IARC Scientific Publication No. 160)*. Lyon: IARC (WHO) and IACR.

- Silcocks, P. B. S., & Robinson, D. 2004. Completeness of ascertainment by cancer registries: putting bounds on the number of missing cases. *J Public Health*, **26**(2), 161–7.
- Silcocks, P. B. S., & Robinson, D. 2007. Simulation modelling to validate the flow method for estimating completeness of case ascertainment by cancer registries. *J Public Health*, **29**(4), 455–62.
- Skeet, R.G. 1991. The hospital-based cancer registry. *Chap. 9, pages 101–107 of:* Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S., & Skeet, R.G. (eds), *Cancer Registration: Principles and methods. (IARC Scientific Publication No. 95)*. Lyon: IARC.
- StataCorp. 2007 (Apr.). *Stata Statistical Software: Release 10*. College Station: Stata Corporation.
- Tilling, K. 2001. Capture-recapture methods—useful or misleading? *Int. J. Epidemiol.*, **30**(1), 12–14.
- Tilling, K., & Sterne, J. 1999. Capture-recapture models including covariate effects. *American Journal of Epidemiology*, **149**(4), 392–400.
- Tingulstad, S., Halvorsen, T., Norstein, J., Hagen, B., & Skjeldestad, F. E. 2002. Completeness and accuracy of registration of ovarian cancer in the cancer registry of Norway. *Int J of Cancer*, **98**(6), 907–911.
- Villard-Mackintosh, L., Coleman, M. P., & Vessey, M. P. 1988. The completeness of cancer registration in England: an assessment from the Oxford-FPA contraceptive study. *Br J Cancer*, **58**(4), 507–11.
- Wagner, G. 1991. History of cancer registration. *Chap. 2, pages 3–6 of:* Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S., & Skeet, R.G. (eds), *Cancer Registration: Principles and methods. (IARC Scientific Publication No. 95)*. Lyon: IARC.
- WHO. 2009 (February). *Fact Sheet No. 297*.
- Winchester, D., Stewart, A., Phillips, J., & Ward, E. 2010. The National Cancer Data Base: Past, Present, and Future. *Annals of Surgical Oncology*, **17**, 4–7. 10.1245/s10434-009-0771-3.

Young, J.L. 1991. The hospital-based cancer registry. *Chap. 13, pages 177–184 of:* Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S., & Skeet, R.G. (eds), *Cancer Registration: Principles and methods. (IARC Scientific Publication No. 95).* Lyon: IARC.