



Quality of Cancer Data and Survival Estimates: A Review of Literature

Houssein N *

Department of Epidemiology, Faculty of Public Health, University of Benghazi, Libya

*Corresponding author: Naeima Houssein, Department of Epidemiology, Faculty of Public Health, University of Benghazi, Libya

Received: 01 August 2024

Accepted: 22 August 2024

Published: 26 August 2024

© 2024 The Author. This is an open-access article and is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Abstract

Cancer registries play a central role in the documentation of cancer information. Over the past few decades, the data quality of cancer registries has been the subject of considerable debate among healthcare planners and has received increasing attention among epidemiologists. This is because of the importance of cancer registries' data in health services planning, and epidemiological research. Cancer registries are responsible for collecting the basic demographic and disease information of every patient diagnosed with cancer and producing high-quality cancer statistics. The quality of cancer registry data is evaluated using different techniques to improve the registration process, completeness, and accuracy. This review aims to describe the quality of cancer registration as reported in the literature, highlighting the effect of the completeness and accuracy of cancer data on survival estimates. A limited number of studies have looked at the quality of cancer data. The existing literature indicated several limitations on the quality of cancer data that influence the estimates of cancer survival and contribute to international variations of cancer survival between countries. This effect could make survival estimates either underestimated or overestimated. No specific data field was reported to be responsible for the change in survival estimate. However, the importance of some clinical fields, such as clinical stages and treatments, has been highlighted in pieces of literature. Survival statistics based on cancer registries were also affected by the presence of death certificate only (DCO) registrations. Complete and accurate data are crucial for obtaining reliable results and valid inferences in oncology research.

Keywords: accuracy, completeness, cancer, data, survival

Abbreviations: SCR: Scottish Cancer Registry, DCO: death certificate only, OCR: Ontario Cancer Registry, EMR: electronic medical record

1. Introduction

Cancer registries play a central role in the documentation of cancer information. Over the past few decades, the data quality of cancer registries has been the subject of considerable debate among healthcare planners and has received increasing attention among epidemiologists. This is because of the importance of cancer registry data in health services planning, epidemiological research, and evaluation of cancer health care services [1]. Data quality at cancer registries is determined by at least four important factors: comparability, completeness, validity (accuracy), and timelines. Each of these prerequisite characteristics has special assessment procedures [2]. For instance, completeness can be evaluated by either qualitative or quantitative methods.

The former includes historic data methods, mortality incidence ratios, and histological verification of diagnosis. The latter examines the degree of completeness of registration by applying independent case ascertainment comparisons, capture-recapture methods, and death certificate methods [3].

Completeness is the extent to which all appropriate data items within the jurisdiction covered by the cancer registry are recorded [4]. This metric plays a fundamental role in assessing the quality of data. Various studies have documented completeness rates that range from 36% to 99.3% [5]. Factors that influence completeness encompass the application of multiple data sources, proactive methods for identifying cases, and the promptness of data submission. Ensuring the thoroughness of case identification by reporting facilities is an essential element of the quality assurance procedures of a

central cancer registry. It is imperative for registries to validate that all relevant tumor cases are being reported by the facilities [6].

Another important dimension of data quality is accuracy, defined as the proportion of cases with specific characteristics that actually have that attribute [4]. Accuracy constitutes a crucial aspect of data quality. The evaluation of accuracy can be accomplished through the comparison of registry data against a reference standard, such as medical records. Studies have documented agreement rates ranging from 84.7% to 99.6% for various data elements [7]. Furthermore, certain registries have pinpointed misclassification in aspects like diagnostic criteria, laterality, and gender as areas necessitating enhancement. The utilization of multiple data resources and logical validations can contribute to the enhancement of accuracy [8].

The quality of cancer data is a crucial aspect of cancer research and patient care. Therefore, ensuring the accuracy and reliability of these data is essential for making informed decisions about treatment options, predicting patient outcomes, and developing effective strategies for cancer prevention and management. Given the critical role of completeness and accuracy of cancer data that has been reported by several researchers, the reviewed studies were concerned with those two aspects of data quality to identify their levels at cancer registries and medical databases. This review aims to describe and discuss the quality of cancer registration as reported in the literature, highlighting the effect of the completeness and accuracy of cancer data on survival estimates.

2. Completeness and Accuracy of Cancer Registries

Previous studies called 'in-house' studies were conducted by Brewster and colleagues [9–15] and published in the period between 1994 and 2008. They worked on the assessment of data accuracy and case ascertainment for all cancer registrations at the Scottish Cancer Registry (SCR), which is responsible for collecting cancer data in Scotland. Some studies have concluded that data at the SCR were found to be at a high level of accuracy and completeness. Furthermore, the overall completeness and accuracy in 1992 at the SCR were estimated at 96.5% [13, 14].

However, accurate capture of pathologic details was challenging. Brewster et al. [11] assessed the accuracy of lung cancer registration using a random sample of 340 patients registered with lung carcinoma in 1990. These registrations were compared with their relevant medical records, which were available for 309 included in the study. Death certificate registrations accounted for 20 cases. The comparison was based on selected data items, including demographic data, treatment and histological verification data, and site and morphologic data. Their results revealed some discrepancies, that were higher for site codes,

morphology codes, and histological verification. These discrepancies were reported as 56.5%, 47.25%, and 12.5%, respectively. The researchers explained that these errors in data items were due to the missed data and the time of re-abstraction of the database for this study.

A similar methodology was undertaken and close findings were reported in other previous studies concerned with colorectal and breast tumors, and non-melanoma skin cancer. Although the overall data quality at the SCR was considered high, there were a few limitations in some data items related to those cancers. This was in addition to the reported under-ascertainment of intracranial tumors since the SCR was missing 46% of cases [16].

In addition, a retrospective review of the SCR data in 1997 found significant issues with reliability for grades of differentiation, staging variables, and dates of treatment [13]. Similar problems have been reported in the Thames Cancer Registry [17]. Accordingly, Klassen et al. [18] pointed out that the stage of disease at diagnosis and histological grade are the two clinical characteristics often missing in the cancer registry.

In a comparison of registry data with hospital data, some authors argue that hospital data would be more reliable than cancer registries [19]. However, Schouten et al. [20] suggested that data collected by clinicians cannot be considered a golden standard' in cancer registries. This is because cancer registries and clinicians may collect data with a 'different perspective'. Data collected by clinicians is mainly used to determine the treatment and prognosis of patients, therefore, it is characterized by less basic detail. Whereas, people who work at cancer registries are trained to follow coding rules. This study, however, was cross-sectional, it measured the difference in the quality of clinical and registry data at one point in time only.

These suggestions have been observed to a certain extent in Gregor et al. study [21], they reported considerable limitations in the medical records available for 91.2% of the participants (4465) diagnosed in 1995. The missing data is noticed in important prognostic fields. For example, 423 cases did not have their staging detail, and 999 patients did not have their microscopic verification. Comparing hospital data with cancer registry was reported in further studies [22, 23].

Regarding other literature on the UK cancer registries, a study found cancer records characterized by high standards of completeness, accuracy, and reliability [24]. However, another study [25] found that cancer data in the United Kingdom has many aspects of incompleteness in many data fields, such as sex, age, staging information, and treatment. In addition, Adams et al. [26] suggest that there is a socioeconomic gradient in the quality of data from cancer registries in the United Kingdom. For example,

death certificate only (DCO) registrations represent the only evidence for a diagnosis of cancer, are the most common available records for the most deprived population. Jones et al. [27] added that such registrations are associated with males rather than females. The presence of these registrations, therefore, reflects the fact that the diagnosis of cancer was post-mortem. In such cases, the date of incidence is the date of death, and this negatively affects the overall cancer survival estimate. Despite the continuing efforts of cancer registries to collect cancer incidence data, some literature has reported evidence of missing and inaccurate cancer registry data under the terms of a voluntary case reporting policy. Consequently, these limitations create frequent problems in data analysis and interpretation. Cancer survival estimates may be under or overestimated based on the processes of data collection, reporting, and analysis.

3. Data Quality and Cancer Survival Estimates

The incompleteness of medical records may introduce a selective bias in the cancer survival estimates of patients with relatively poor or good prognoses. Brenner et al. [28] examined the impact of data incompleteness on five-year survival estimates in patients aged 15 years and older, with a first diagnosis of one of the common types of cancer in Finland, in the period between 1990 and 1999. They concluded that selective under-ascertainment of patients with a good prognosis may lead to an underestimation of cancer patient survival, whereas an opposite effect could result from selective under-ascertainment of patients with poor prognosis.

A study from Germany reported some incompleteness of Hamburg Cancer Registry data on certain malignancies such as colon, prostate, and urinary bladder. The authors conclude that such limitations in data quality may impact the validity of the cancer registry to produce convincing survival estimates [29].

Survival estimates based on data from cancer registries are also affected by the presence of DCO registrations: in addition to the incompleteness of case ascertainment [30, 31]. DCO cases are associated with patients who have shorter survival times on average. Pollock et al. [30] examined the impact of adding DCO registrations to district health authority data on the estimation of 5-year survival of colorectal cancer. As mentioned earlier, for DCO registrations, the incidence date (date of diagnosis) and date of death are the same. Therefore, the duration of survival is considered to be zero. They found that survival rates were decreased by 8.6%. They concluded that the DCO registrations play an important role in declining the survival estimate for a concerned population, and the exclusion of DCO is not the solution to the issue, but rather the improvement of the quality of DCO. Likewise, another study examined the relationship

between those two factors and pointed out that the percentage of high DCO indicates a poor case ascertainment. However, low DCO registrations do not reflect a complete case ascertainment [32].

Moreover, Berrino [33] suggested that these factors may cause either under or over-cancer survival estimations. A study by Robinson et al. [34] aimed to assess the impact of DCO registrations and incomplete ascertainment on survival estimates of certain cancer sites, including lung cancer. Their methodology was based on comparing 5-year survival estimates using data from the Finland and Thames cancer registries. They compared survival estimates before and after adjustment for DCO registrations, incompleteness, and both. Their findings confirmed the observations that were reported by Berrino [33] and have shown that the 5-year survival estimate was influenced by the presence of both DCO and under ascertainment, they reported that Finland registry data are visually complete and had few DCO registrations due to the obligatory legislation of cancer registration in Finland. Consequently, adjustments for those factors had little effect on survival estimates.

On the other hand, the Thames Cancer Registry in the United Kingdom, where cancer registration is voluntary, had data showing less completeness and higher DCO proportions. As a result, adjusting for under-ascertainment caused a substantial increase in survival estimates, whereas, adjusting for DCO led to a marked reduction in survival estimates. In other words, the presence of DCO and incomplete case details in cancer registries can cause bias in the calculated survival estimate in opposite directions. This study relied on the proportions of DCO both in Finland and Thames cancer registries as a measure for the quality of data. This approach, however, was criticized as only providing a broader indicator of data quality and not accurately estimating the completeness [32]. In addition, it would not be sufficient to reflect the overall completeness of data in both registers as the researchers assessed the completeness for sex and cancer site fields only.

Beral et al. [35] have reported that the incompleteness in the UK cancer registrations made survival estimates misleading and seem significantly worse than they are. In contrast, mortality statistics are more reliable than survival rates. This is because death certificates are legally required for burial and cremation. Therefore, death information in the cancer registry is almost complete, as the death certificates are automatically transferred to the regional registries. The authors discussed many limited aspects of the UK cancer survival statistics which are based on cancer registries' data. They suggest that the absence of non-fatal case registrations can distort the overall survival estimate. In addition, cancer screening also causes substantial distortions by artificially prolonging the recorded survival duration. Similarly, another recent British study reported that incompleteness of case ascertainment can lead to bias

in estimates of cancer survival producing very low survival figures [36].

Different findings have been reported where registration is mandatory. In Denmark, Norgaard et al. [22] examined the quality of data on hematological malignancies in a population-based hospital discharge registry by independent comparison with data from the Danish Cancer Registry. Further, they investigated the impact of missing data on survival estimates. They used the positive predictive value (PPV) measure to assess the degree of completeness. This approach indicates that most positive test results are true positives, which can be useful in assessing the quality of the registry data. As a result, they found similarities in the Kaplan-Meier survival curves and the completeness level.

The findings of this study were consistent with Hall et al. study [4] that assessed the quality of the Ontario Cancer Registry (OCR) data. They had clinical information on 898 patients with squamous carcinoma of the head and neck, including index tumor site, date of diagnosis, vital status, date of death, and cause of death from a prospective database at the Kingston Regional Cancer Centre, compared to the same data elements in the OCR for the same patients. Their results indicate a high level of case ascertainment at the registry and no differences in the survival estimate between the two data sources.

Completeness of ascertainment and accuracy of registration with other factors significantly contribute to the international variations in cancer survival [37]. These variations have been widely reported in the literature. More importantly, the EURO CARE publications have presented strong evidence of lung cancer survival differences between the European countries. In particular, the EURO CARE-4 study indicates the drawbacks of under-reporting of cancer data on the accuracy of international cancer survival comparison and interpretation; consequently, survival estimates could not be reliable in evaluating the quality of cancer care [38].

The quality of cancer data and its impact on survival estimates are influenced by various factors outlined in recent research. Issues such as loss to follow-up cases can lead to biased survival estimates, especially in small population-based cancer registries, with censoring of cases potentially overestimating survival rates [39]. Accurate survival estimates require the combination of high-quality cancer registry data with electronic medical record (EMR) data. The use of EMR data alone for survival estimates can result in an overestimation of survival times [40]. Additionally, missed deaths within cancer registry data can result in inflated long-term survival estimates, highlighting the importance of detecting and addressing such discrepancies [41].

More importantly, survival rates are considered essential indicators of the effectiveness of cancer services, reflecting the prospects of cure and the

overall quality of cancer care [42]. Several studies underscore the critical role of high-quality cancer data for measuring survival and emphasize the need for accurate and comparable survival estimates to assess the overall effectiveness of cancer care in a population, guiding treatment decisions, and improving outcomes for cancer patients [43–45].

In summary, cancer registries collect the basic demographic and disease information of every patient diagnosed with cancer. This routinely collected information produces high-quality research about the incidence, patterns, and mortality from cancer-related conditions. Incomplete or inaccurate data can lead to misinformation about cancer statistics and services. The quality of cancer registry data is evaluated using different techniques to improve the registration process, completeness, and accuracy.

The review aimed to shed light on the existing literature regarding the quality of cancer data at registries and medical settings in terms of completeness and accuracy, and how these aspects affect cancer survival estimates. This review was limited to certain literature sources including Ovid MEDLINE in process, EMBASE, PubMed, and Web of Knowledge from inception until 2023. Further relevant studies were identified by searching references in the obtained full-text papers. However, only articles published in English were included. Being a descriptive review, it does not systematically search and review the completeness and accuracy of cancer registration and its effect on cancer survival estimates. However, it tried to bridge the gap in knowledge and practice regarding the significance of data quality in cancer statistics.

4. Conclusion

A limited number of studies have looked at the quality of cancer data. The existing literature indicated several limitations on the quality of cancer data that influence the estimates of cancer survival and contribute to international variations of cancer survival between countries. This effect could make survival estimates either underestimated or overestimated. No specific data field was reported to be responsible for the change in survival estimate. However, the importance of some clinical fields such as clinical stage and treatments has been highlighted in pieces of research. The presence of DCO registrations can significantly impact the accuracy and interpretation of cancer survival estimates from registry data, and appropriate measures should be taken to address this issue. Furthermore, maintaining high levels of completeness and accuracy is crucial for cancer registries to provide reliable data for cancer control planning, research, and policy development. Standardized metrics and reporting of data quality are necessary to ensure confidence in the usefulness of cancer registry data.

References

1. Parkin DM, Chen VW, Ferlay J, et al. Comparability and Quality Control in Cancer Registration. ARC Technical Report.1994. No. 19.
2. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *Eur J Cancer*. 2009;45(5):747-55.
3. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. *Eur J Cancer*. 2009;45(5):756-64.
4. Hall S, Schulze K, Groome P, et al. Using cancer registry data for survival studies: the example of the Ontario Cancer Registry. *J Clin Epidemiol*. 2006;59(1):67-76.
5. Giusti F, Martos C, Negrão Carvalho R, et al. Quality indicators: completeness, validity and timeliness of cancer registry data contributing to the European Cancer Information System. *Front Oncol*. 2023;13:1219128.
6. Shokrizadeharani L, Batooli Z, Heydarian S, et al. Evaluation of Completeness, Comparability, Validity, and Timeliness in Cancer Registries: A Scoping Review. *Stud Health Technol Inform*. 2023;305:160-63.
7. Sirirungreung A, Buasom R, Jiraphongsa C, et al. Data Reliability and Coding Completeness of Cancer Registry Information Using Reabstracting Method in the National Cancer Institute: Thailand, 2012 to 2014. *J Glob Oncol*. 2018;4:1-9.
8. Kalsi JK, Ryan A, Gentry-Maharaj A, et al. Completeness and accuracy of national cancer and death registration for outcome ascertainment in trials-an ovarian cancer exemplar. *Trials*. 2021;22(1):88.
9. Brewster D, Crichton J, Muir C. How accurate are Scottish cancer registration data? *Br J Cancer*. 1994;70(5):954-59.
10. Brewster D, Muir C, Crichton J. Registration of colorectal cancer in Scotland: an assessment of data accuracy based on review of medical records. *Public Health*. 1995;109(4):285-92.
11. Brewster D, Muir C, Crichton J. Registration of lung cancer in Scotland: an assessment of data accuracy based on review of medical records. *Cancer Causes Control*. 1995;6(4):303-10.
12. Brewster DH, Crichton J, Harvey JC, et al. Benefits and limitations of pathology databases to cancer registries. *J Clin Pathol*. 1996;49(11):947-49.
13. Brewster DH, Crichton J, Harvey JC, et al. Completeness of case ascertainment in a Scottish regional cancer registry for the year 1992. *Public Health*. 1997;111(5):339-43.
14. Brewster DH, Stockton D, Harvey J, et al. Reliability of cancer registration data in Scotland, 1997. *Eur J Cancer*. 2002;38(3):414-17.
15. Brewster DH, Stockton DL. Ascertainment of breast cancer by the Scottish Cancer Registry: an assessment based on comparison with five independent breast cancer trials databases. *Breast*. 2008;17(1):104-06.
16. Counsell CE, Collie DA, Grant R. Limitations of using a cancer registry to identify incident primary intracranial tumours. *J Neurol Neurosurg Psychiatry*. 1997;63(1):94-97.
17. Pollock AM, Vickers N. Reliability of data of the Thames cancer registry on 673 cases of colorectal cancer: effect of the registration process. *Qual Health Care*. 1995;4(3):184-89.
18. Klassen AC, Curriero F, Kulldorff M, et al. Missing stage and grade in Maryland prostate cancer surveillance data, 1992-1997. *Am J Prev Med*. 2006;30(2 Suppl):S77-87.
19. Nwene U, Smith A. Assessing completeness of cancer registration in the north-western region of England by a method of independent comparison. *Br J Cancer*. 1982;46(4):635-39.
20. Schouten LJ, Jager JJ, van den Brandt PA. Quality of cancer registry data: a comparison of data provided by clinicians with those of registration personnel. *Br J Cancer*. 1993;68(5):974-77.
21. Gregor A, Thomson CS, Brewster DH, et al. Management and survival of patients with lung cancer in Scotland diagnosed in 1995: results of a national population based study. *Thorax*. 2001;56(3):212-17.
22. Nørgaard M, Skriver MV, Gregersen H, et al. The data quality of haematological malignancy ICD-10 diagnoses in a population-based hospital discharge registry. *Eur J Cancer Prev*. 2005;14(3):201-06.

23. Field K, Kosmider S, Johns J, et al. Linking data from hospital and cancer registry databases: should this be standard practice? *Intern Med J.* 2010;40(8):566-73.
24. Hogget C. Nationwide audit of the quality of data held by regional cancer registries. 1995. Bristol: University of Bristol.
25. Warren JL, Harlan LC. Can cancer registry data be used to study cancer treatment? *Med Care.* 2003;41(9):1003-005.
26. Adams J, Audisio RA, White M, et al. Age-related variations in progression of cancer at diagnosis and completeness of cancer registry data. *Surgical Oncology.* 2004;13(4):175-79.
27. Jones AP, Haynes R, Sauerzapf V, et al. Geographical access to healthcare in Northern England and post-mortem diagnosis of cancer. *J Public Health (Oxf).* 2010;32(4):532-37.
28. Brenner H, Hakulinen T. Population-based monitoring of cancer patient survival in situations with imperfect completeness of cancer registration. *Br J Cancer.* 2005;92(3):576-79.
29. Nennecke AL, Hentschel S, Reintjes R. Cancer survival analysis in Hamburg 1995-2003: assessing the data quality within a population-based registry. *Acta Oncol.* 2009;48(1):34-43.
30. Pollock AM, Vickers N. The impact on colorectal cancer survival of cases registered by 'death certificate only': implications for national survival rates. *Br J Cancer.* 1994;70(6):1229-231.
31. Stotter A, Bright N, Silcocks PB, et al. Effect of improved data collection on breast cancer incidence and survival: reconciliation of a registry with a clinical database. *BMJ.* 2000;321(7255):214.
32. Brenner H. Limitations of the death certificate only index as a measure of incompleteness of cancer registration. *Br J Cancer.* 1995;72(2):506-10.
33. Berrino F. The EUROCARE Study: strengths, limitations and perspectives of population-based, comparative survival studies. *Ann Oncol.* 2003;14 Suppl 5:v9-13.
34. Robinson D, Sankila R, Hakulinen T, et al. Interpreting international comparisons of cancer survival: the effects of incomplete registration and the presence of death certificate only cases on survival estimates. *Eur J Cancer.* 2007;43(5):909-13.
35. Beral V, Peto R. UK cancer survival statistics. *BMJ.* 2010;341:c4112.
36. Møller H, Richards S, Hanchett N, et al. Completeness of case ascertainment and survival time error in English cancer registries: impact on 1-year survival estimates. *Br J Cancer.* 2011;105(1):170-76.
37. Black RJ, Sankaranarayanan R, Parkin DM. Interpretation of population-based cancer survival data. *JIRC Sci Puli.* 1998; 145, pp.7-13.
38. Berrino F, De Angelis R, Sant M, et al. Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EURO CARE-4 study. *Lancet Oncol.* 2007;8(9):773-83.
39. Gil F, Miranda-Filho A, Uribe-Perez C, et al. Impact of the Management and Proportion of Lost to Follow-Up Cases on Cancer Survival Estimates for Small Population-Based Cancer Registries. *J Cancer Epidemiol.* 2022;2022:9068214.
40. Gensheimer MF, Narasimhan B, Henry AS, et al. Accuracy of Electronic Medical Record Follow-Up Data for Estimating the Survival Time of Patients With Cancer. *JCO Clin Cancer Inform.* 2022;6:e2200019.
41. Dahm S, Barnes B, Kraywinkel K. Detection of missed deaths in cancer registry data to reduce bias in long-term survival estimation. *Front Oncol.* 2023;13:1088657.
42. Shulman LN, Palis BE, McCabe R, et al. Survival As a Quality Metric of Cancer Care: Use of the National Cancer Data Base to Assess Hospital Performance. *J Oncol Pract.* 2018;14(1):e59-e72.
43. Li R, Abela L, Moore J, et al. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiol.* 2014;38(3):314-20.
44. Loya H, Poduval P, Anand D, et al. Uncertainty Estimation in Cancer Survival Prediction. *arXiv.*2020.
45. Plana D, Fell G, Alexander BM, et al. Cancer patient survival can be parametrized to improve trial precision and reveal time-dependent therapeutic effects. *Nat Commun.* 2022;13(1):873.

To access the full-text version of this article,
please scan the QR code:

