HHS Public Access

Author manuscript

Lung Cancer. Author manuscript; available in PMC 2020 August 01.

Published in final edited form as:

Lung Cancer. 2019 August ; 134: 16–24. doi:10.1016/j.lungcan.2019.05.016.

Misclassification of the actual causes of death and its impact on analysis: A case study in non-small cell lung cancer

Kay See Tan, PhD

Department of Biostatistics and Epidemiology Memorial Sloan Kettering Cancer Center 485 Lexington Ave, 2nd Floor New York, NY 10017 United States

Abstract

Objectives: Cumulative incidence of lung cancer deaths (LC-CID) is an important metric to understand cancer prognosis and to determine treatment options. However, credible estimates of LC-CID rely on accurate cause-of-death coding in death certificates. Results from lung cancer screening trials estimated 15% under-reporting and 1% over-reporting of lung cancer deaths due to misclassification. This study investigated the impact of cause-of-death misclassification on the estimation of LC-CID.

Materials and Methods: Patients with stage I/II non-small cell lung cancer (NSCLC) from the Surveillance, Epidemiology, and End Results registry were included. LC-CID was estimated using the competing-risk approach in two ways: (1) reporting observed estimates that ignore potential cause-of-death misclassification and (2) correcting for plausible misclassification rates reported in the literature (15% under-reporting and 1% over-reporting). Bias was quantified as the difference between observed and corrected 10-year LC-CIDs: positive values indicated that observed LC-CID overestimated true LC-CID, whereas negative values indicated the opposite.

Results: Among 66,179 patients, the impact of over-reporting on 10-year LC-CID was negligible across all age groups. In contrast, under-reporting resulted in substantial underestimation of 10-year LC-CID. The biases increased as age increased due to higher LC-CIDs: 10-year LC-CIDs among stage I patients 18–44, 45–59, 60–74 and ≥75 years were 25%, 32%, 41%, and 50%, respectively, and the corresponding biases given the plausible misclassification rates were −4.4%, −5.6%, −7.1%, and −8.6%. Because the observed LC-CIDs among patients with stage II disease were higher than those with stage I disease, the biases were greater among stage II patients, up to −12.5% in the oldest age group.

*Corresponding author: Kay See Tan, PhD, Department of Biostatistics and Epidemiology, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, 2nd Floor, New York, NY 10017, United States tank@mskcc.org.

Author contributions: KST conceived the present work, analyzed the data, and wrote the manuscript. KST extracted all the data in the study and takes full responsibility for the integrity of the data and the accuracy of the data analysis.

Summary Conflicts of interest statement: The author has no potential conflicts of interest to disclose.

Additional information: The supplemental tables, figures and R codes to generate results in this study can be found in the Supplemental Materials.

Declarations of interest: none

Conclusions: In lung cancer, LC-CID may be severely underestimated due to under-reporting of lung cancer deaths, particularly among older patients or those with late-stage disease. Future studies that involve such subpopulations should present the corrected LC-CIDs based on plausible misclassification rates alongside the observed LC-CIDs.

Keywords

Competing risk events; cause of failure; cumulative incidence; death certificate; cause-specific survival; Epidemiology and Prevention

1. Introduction

Accurate estimation of long-term survival is crucial to the understanding of disease prognosis and cancer-specific mortality. However, what complicates the compilation of this data is the possibility of patients dying from causes unrelated to cancer. Patient with lung cancer, for example, die from either lung cancer or non-lung cancer causes. In these instances, when the occurrence of one outcome precludes the occurrence of the other, these two outcomes are considered competing events [1, 2]. In the presence of competing events, epidemiologic studies commonly report cancer-specific mortality as the cumulative incidence of cancer-specific deaths.

Cumulative incidence of cancer-specific deaths is a survival measure that represents *actual prognosis*, which acknowledges that the probability of dying from cancer may be countered by the probability of dying from other competing causes [3–5]. This contrasts with *cancer prognosis* (or net survival), which assumes cancer as the only possible cause of death and ignores competing causes.

When determining the cumulative incidence of cancer-specific deaths in a population, it is necessary to identify whether the patients died of cancer-related causes. However, obtaining the actual causes of death are often quite difficult. Death certificates commonly include incomplete or incorrect information relating to cause of death, due to diagnostic and coding errors [6, 7]. In the context of lung cancer, cause-of-death misclassification occurs when lung cancer deaths are misclassified as non-lung cancer deaths, or when non-lung cancer deaths are misclassified as lung cancer deaths.

The accuracy of cause-of-death information abstracted from death certificates has been examined in detail across major cancer sites [8–14]. Many of these studies confirmed the reliability of death certificates among cancer sites such as breast, prostate, and pancreas [11, 12], while others reported high misclassification rates, particularly among colon, liver, and oral cancers, and strongly advised against relying on death certificates alone for cause-of-death assignments [9, 12, 15]. In fact, several authors demonstrated that misclassification rates could be as high as 15% among lung cancer cases, and the impact of misclassification on cancer prognosis measures may be substantial [8, 16, 17]. However, these studies isolated the effect of cancer on survival by removing competing causes of death rather than accounting for competing outcomes to estimate “real world” probabilities of cancer-specific deaths -- the information most relevant to clinicians and patients. To date, only two studies [18, 19] systematically evaluated the impact of misclassification on actual prognosis using

the competing risk framework. Both reported that cause-of-death misclassification resulted in inaccurate estimation of cumulative incidence of breast cancer deaths. As the composition of cancer to non-cancer deaths are different between breast cancer and lung cancer [3], however, the conclusions from these two studies may not be applicable to lung cancer.

Given these limitations, the present study investigates the impact of cause-of-death misclassification on the estimation of cumulative incidence of lung cancer deaths (LC-CID). This study is the first to examine cause-of-death misclassification in the competing risk framework among patients with early-stage non-small cell lung cancer (NSCLC).

2. Materials and Methods

2.1. Study Population

Patients diagnosed between 2004 and 2014 with NSCLC were included from the Surveillance, Epidemiology, and End Results (SEER) registry released in April 2017 [20] (See flow diagram, e-Figure 1). In the SEER registry, the cause-specific death classification variable uses two different rules to assign the cause of death depending on whether the index tumor was an individual's first and only tumor or the first of multiple tumors. This distinction addresses the evidence of greater ambiguity in cause-of-death assignments among patients diagnosed with more than one cancer [12, 21]. Therefore, to maintain consistency in the definition of cancer-specific deaths, patients with prior cancers or second primaries were excluded from the analysis cohort. Histological diagnoses of NSCLC were based on the International Classification of Diseases for Oncology codes (ICD-O-3, e-Appendix A) [22]. For each patient, year of diagnosis, stage (based on the sixth edition of the American Joint Committee on Cancer (AJCC) Staging Manual), histology, age, sex, surgery (yes/no), vital status (alive/dead), cause of death (lung cancer-specific or non-lung cancer-specific), and survival duration (from diagnosis to death or last follow-up) were recorded. The analysis cohort comprised of NSCLC patients aged ≥18 years with stage I or II diseases at the time of diagnosis.

2.2. Cumulative incidence and competing risks

In the setting of NSCLC, deaths are attributed to lung cancer (the events of interest) or non-lung cancer causes (the competing events). Conventionally, the cumulative incidence of deaths can be obtained using the Kaplan-Meier approach (1 minus Kaplan-Meier estimate). The Kaplan-Meier approach provides a nonparametric estimate of the overall survival probability of an event of interest [1]. However, in the presence of non-lung cancer deaths, the Kaplan-Meier approach may not be appropriate to estimate the cumulative incidence of lung cancer deaths because it assumes patients will not die from other causes [1, 2]. Instead, the cumulative incidence of lung cancer deaths can be examined using competing-risk analysis to account for the presence of competing events. Rather than censoring the individuals who died from non-lung cancer causes (as done in the Kaplan-Meier approach), the competing risk framework accounts for non-lung cancer deaths (competing events) in the analysis to estimate LC-CID. In each time interval of the cumulative incidence function, the incidence of lung cancer deaths is estimated as the joint probability of experiencing lung cancer deaths in this time interval given that the individuals survived both lung cancer and

non-lung cancer deaths in all prior intervals [1]. Even so, the cumulative incidence estimates are still susceptible to cause-of-death misclassification.

2.3. Types of cause-of-death misclassification

This paper applies the formal definitions of misclassification described by Bakoyannis and Yiannoutsos [19]. Additional details to accompany the brief descriptions below can be found in the appendix (e-Appendix B). With two competing outcomes, there may be misclassification of the events of interest and/or the competing events (Figure 1). Probabilities are assigned to each of these scenarios, ranging from 0 to 1.

Let ***a*** = the probability of misclassification of the events of interest (e.g., lung cancer deaths misreported in death certificates as non-lung cancer deaths and thus "under-reporting" lung cancer deaths).

Let ***b*** = the probability of misclassification of the competing events (e.g., non-lung cancer deaths misreported as lung cancer deaths and thus "over-reporting" lung cancer deaths).

When a=0 and b=0, there is no misclassification error. This scenario is ideal but unlikely. In reality, cause-of-death recording in death certificates are subject to errors, broadly categorized into two types: (1) unidirectional misclassification and (2) bidirectional misclassification. Unidirectional misclassification occurs when only one of the two outcomes are misclassified (either a=0 with b>0, or b=0 with a>0). Bidirectional misclassification occurs when both the events of interest and competing events are misclassified (both a>0 and b>0). Bidirectional misclassification applies to our current focus on the estimation of lung cancer-specific mortality.

Because a and b are probabilities, the influence of a and b are relative to the magnitude of LC-CID and the corresponding cumulative incidence of non-lung cancer deaths (non-LC-CID). The higher the LC-CID, the greater the impact of a; similarly, the higher the non-LC-CID, the greater the impact of b. For example, if the true proportion of patients who died from lung cancer was 50% and the degree of under-reporting, a, was 10%, then 5% of the *total* population would have been misclassified as non-lung cancer deaths while having lung cancer deaths. But if the true proportion of patients who died from lung cancer causes was lowered to 10% while a remains at 10%, under-reporting would only impact 1% of the total population.

2.4. Statistical Analysis

Patients were categorized into eight subgroups based on stage (I or II) and age at diagnosis (18 to 44, 45 to 59, 60 to 74, and ≥75 years). In each subgroup, the observed LC-CID and non-LC-CID were estimated. The observed estimate of LC-CID reports the observed incidence under the assumption that cause-of-death coding is accurate. In reality, the observed LC-CID estimates are susceptible to biases due to cause-of-death misclassification. The observed LC-CID can be corrected to derive the *true* LC-CID if the actual probabilities of misclassification are known [19]. Given a and b, the corrected LC-CID can be derived as (see details in e-Appendix B):

$$\text{corrected LC-CID} = \frac{1-b}{1-a-b}(\text{observed LC-CID}) - \frac{b}{1-a-b}(\text{observed non-LC-CID}).$$

In this study, like most observational studies, the exact misclassification probabilities were unknown. Upon literature review, plausible estimates of a and b were identified from two National Cancer Institute sponsored lung cancer screening trials conducted in the United States [23]. Both the Mayo Lung Project [16, 24, 25] and the Johns Hopkins Lung Project [26] conducted formal mortality reviews for all deaths by two or more experts and estimated the accuracy of cause-of-death information reported in death certificates. In the Mayo Lung Project, death certificates misclassified 11.4% (27/297) of lung cancer deaths as non-lung cancer deaths and misclassified 0.9% (14/1636) of non-lung cancer deaths as lung cancer deaths (i.e., a = 11.4% and b = 0.9%). The corresponding estimates in the Johns Hopkins Lung Project were 15.5% (51/329) and 1.1% (13/1203; i.e., a = 15.5% and b = 1.1%). Other validation studies reported similar values of a: for example, a prospective cohort study among workers exposed to asbestos reported that 14% of autopsy-confirmed lung cancer deaths were misclassified as non-lung cancer deaths in death certificates [27].

Based on these studies, plausible values of the misclassification rates were reliably estimated as a = 15% (moderate under-reporting) and b = 1% (negligible over-reporting). The impact of these misclassification rates on LC-CID estimates were quantified in each subgroup. To generalize the relationships between different misclassification rates and LC-CID estimates, the biases in LC-CID were also assessed under additional scenarios: a = 0% (no under-reporting) or 5% and b = 0% (no over-reporting) or 5%.

Hence, in each scenario, LC-CIDs were estimated using the competing-risk approach in two ways: (1) reporting observed estimates that ignore potential cause-of-death misclassification and (2) correcting for misclassification rates a and b using the correction method described above. Biases were quantified as observed minus corrected 10-year LC-CIDs. A positive value of bias indicates that the observed LC-CID overestimated the true LC-CID, whereas a negative value of bias indicates that the observed LC-CID underestimated the true LC-CID.

As an exploratory assessment, the analyses were repeated in the subcohort of patients who underwent surgery. All analyses were conducted using R 3.5.3 (R Core Team, Vienna, Austria): the *SEERaBomb* package for data processing and the *cmprsk* package for estimation of LC-CID and non-LC-CID. R code to perform this analysis is available in e-Appendix C.

3. Results

3.1. Characteristics and Cumulative Incidence of Death

Analyses included 66,179 patients with stage I (80%) or II NSCLC, of whom 82% were aged ≥60 years, and 50% were females (Table 1). The proportions of lung cancer deaths were higher than those of non-lung cancer deaths across all subgroups. Overall, 10-year LC-CID was 47%, compared to non-LC-CID of 25%, without accounting for cause-of-death misclassification. Ten-year LC-CID increased with advancing age and stage, from 25% to

50% among stage I patients and from 48% to 71% among stage II patients between the youngest and oldest age groups (Figure 2). In contrast, non-LC-CID actually decreased with higher stage across all age groups (e.g., 10-year non-LC-CID among those ≥75 years was 36% among stage I and 23% among stage II). Even though the total number of deaths increased with age, the proportion of deaths attributed to cancer in the older age group decreased because of increased competing causes of death (Table 1: e.g., stage I: 79% of all deaths among those 18–44 years were attributable to lung cancer causes, compared to 64% among those ≥75 years).

3.2. Impact of misclassification in one age group

Results are first described among patients aged 60 to 74 years (stage I) before extending the findings to the other subgroups. The observed 10-year LC-CID was 41% in this age group, compared to non-LC-CID of 25% (Table 1). When $a = 0$ (no under-reporting), any degree of over-reporting ($b > 0$) resulted in the overestimation of LC-CIDs (Figure 3A: given $a = 0\%$ and $b = 5\%$, bias of observed 10-year LC-CID from corrected LC-CID = +1.3%). However, under the literature-based plausible value of $b = 1\%$, this overestimation of LC-CID was negligible (bias = +0.2%). Conversely, when $b = 0$ (no over-reporting), any degree of under-reporting ($a > 0$) resulted in the underestimation of LC-CIDs. Given the literature-based plausible value of $a = 15\%$, the underestimation was as high as 7.3% (bias = −7.3%) in this age group. Though not the outcome of interest, the first scenario would lead to underestimation of non-LC-CID, while the second scenario would lead to overestimation of non-LC-CID.

When both types of misclassification were present, the degree of bias depended on the values of a and b and the magnitudes of LC-CID and non-LC-CID. Because 10-year LC-CID was much higher than non-LC-CID in this age group (41% vs 25%) the impact of a (under-reporting) resulted in greater impact on observed LC-CID than b (over-reporting) did when both a and b were the same values (e.g., the vertical distances between the 0% and 5% curves on Figure 3A across increasing values of a were much smaller than those in Figure 3B across increasing values of b).

3.3 Impact of misclassification among patients with stage I disease

Assuming the literature-based plausible value of $b = 1\%$, the impact of b (over-reporting) on LC-CID was negligible across all each age groups with stage I disease regardless of the level of under-reporting, a (Figure 4A–D). In contrast, assuming the plausible value of $a = 15\%$, the impact of a (under-reporting) on LC-CID was more pronounced. Specifically, the degree of underestimation increased as age increased due to higher LC-CIDs (Figure 4E–H). The 10-year LC-CIDs among those 18–44, 45–59, 60–74 and ≥75 years were 25%, 32%, 41%, and 50%, respectively, and the corresponding biases under plausible misclassification rates ($a = 15\%$ and $b = 1\%$) were −4.4%, −5.6%, −7.1%, and −8.6%.

3.4 Impact of misclassification among patients with stage II disease

The LC-CIDs among patients with stage II disease were higher than those with stage I disease (Table 1; Figure 2). Hence, the impact of a on LC-CID was greater among patients with stage II disease within each age group than those with stage I disease (Figure 5). The

10-year LC-CIDs among those 18–44, 45–59, 60–74 and ≥75 years were 48%, 57% 63%, and 71%, respectively, and the corresponding biases under the plausible misclassification rates (a = 15% and b = 1%) were −8.5%, −10.1%, −11.1%, and −12.5%. In contrast, because non-LC-CIDs were lower among those with stage II disease than those with stage I disease (Table 1), the impact of b was smaller among patients with stage II disease. However, given the extremely small plausible value of b = 1%, the biases in observed LC-CIDs were considered negligible across all age groups (Figure 5).

3.5 Impact of misclassification among patients who received surgery

The analyses above were repeated among the subcohort of patients who received surgery (N=45,114; 68.2%). All-cause mortality was lower among surgery patients compared to the whole cohort, mainly due to the reduction in deaths due to lung cancer (e-Table 1; e-Figure 2). Although the non-LC-CIDs among patients who received surgery were similar to those observed in the whole cohort, 10-year LC-CIDs were lower among those who received surgery compared to the whole cohort (37% vs 47%). As a result, the impact of a on the underestimation of observed LC-CIDs in this surgical cohort were smaller than those observed in the whole cohort given the assumed plausible misclassification rates (e-Figure 3, e-Figure 4).

4. Discussion

Cause-of-death misclassification is inevitable when using death certificates to estimate cancer-specific mortality. When determining the cumulative incidence of cancer-specific deaths, the correction method presented in this paper can be easily applied to account for misclassification in the estimation procedure. Whenever possible, the corrected measures should be presented alongside the observed measures as a sensitivity analysis, in conjunction with relevant misclassification rates obtained from the literature (e.g., validation studies). In the absence of reliable misclassification rates, ranges of possible misclassification probabilities can be utilized (e-Appendix D). As demonstrated in this study of NSCLC patients, the impact of over-reporting of lung cancer deaths was negligible due to extremely low levels of misclassification probabilities obtained from the literature. In contrast, the level of under-reporting was approximately 15%, which had the most substantial impact on observed LC-CID in populations with greater LC-CID, such as patients with late-stage disease and older age.

Based on plausible misclassification rates, the observed LC-CID tended to be underestimated because of misclassification in death certificates. This conclusion contrasts starkly with the results from breast cancer patients using the same registry, where the cumulative incidences of breast cancer deaths were overestimated due to misclassification of causes of death [18]. The primary reason lies in the ratio of the events of interest to the competing events. In lung cancer, the cumulative incidence of cancer-specific deaths was higher than other causes (e.g., 10-year LC-CID vs non-LC-CID were 71% vs. 23% among the oldest age group with stage II NSCLC), whereas the opposite pattern was true in breast cancer (e.g., 10-year cumulative incidence of breast-cancer vs. non-breast cancer deaths were 9% vs. 74% among those aged ≥85 years). This difference in conclusions confirmed

our motivation to study the impact of cause-of-death misclassification specifically among lung cancer patients.

The findings of this study confirm that the impact of under-reporting was most substantial among the oldest age groups due to higher LC-CIDs. Additionally, national mortality registries were shown to overestimate disease-specific causes of deaths among the elderly [28]. For future studies that involve such subpopulations, it is crucial to present the corrected LC-CIDs based on plausible misclassification rates alongside the observed LC-CIDs. Alternatively, one may also report *relative survival* among these subpopulations, thus removing the need for cause-of-death information [3]. Relative survival is the ratio of the overall survival among patients diagnosed with lung cancer and the expected survival in a matched general population (i.e., the excess mortality attributable to cancer). The advantage of reporting relative survival is that it does not require any cause-of-death information, hence removing the concern of misclassification. However, the accuracy of relative survival estimates relies on the comparability between the general population and the lung cancer patient population. For example, some have argued that since most patients with lung cancer are smokers, they are not comparable to the general population, in which the majority of patients are more likely to be nonsmokers [29]. Furthermore, relative survival quantifies cancer prognosis, whereas LC-CID measures actual prognosis. Formal comparisons between these two estimation frameworks can be found elsewhere [2, 3, 30, 31]. Researchers should base the choice of measure on the availability and accuracy (whether cause-of-death assignments or comparable matched cohort) of data and the interpretation of interest.

The concepts and correction for potential misclassification described in this study can be extended to other settings with competing outcomes. For example, if the events of interest are cancer recurrences, it is common to present the cumulative incidence of recurrences. However, some patients may have been incorrectly assigned as dead without recurrence (the competing events) instead of true recurrence events if not for errors or lapses in follow-up scans.

In addition to cause-of-death misclassification, details related to the actual causes of death may not be available for all individuals who died; hence deaths may be recorded as unknown (i.e., deaths cannot be ascribed to lung cancer or other causes with certainty). It is possible that some fraction of the unknown deaths may be due to cancer-specific causes. Gamel and Vogel [31] proposed partitioning these deaths with unknown causes according to the ratio of lung cancer deaths versus deaths attributable to other causes. An alternative approach involves imputing missing causes of death [32].

A few limitations apply to this paper. First, the correction method only allows for two potential outcomes (lung cancer or non-lung cancer deaths) even though non-lung cancer deaths can be further categorized as cardiac and other deaths. Second, the misclassification probabilities a and b were assumed to be constant throughout follow-up duration. Accounting for both limitations will require extensive methodological efforts beyond the scope of this paper. Third, this paper focused only on the impact of misclassification on the cumulative incidence of deaths. Other studies have investigated the statistical properties of misclassification on cause-specific hazard functions using log-rank tests [33], semi-

parametric approach [34] and non-parametric approach [35], and proposed a likelihood-based parametric competing risk analysis [36]. Our focus on the cumulative incidence estimation is motivated by its expanding use in epidemiologic studies. Finally, the use of the SEER registry is subject to inherent limitations associated with any large, multi-institutional registry-based data involving heterogenous patients not treated uniformly.

5. Conclusions

Published estimates of cumulative incidence and cancer prognosis inform clinical care and treatment decisions of lung cancer patients. Credible estimates of the cumulative incidence of lung cancer deaths rely on accurate cause-of-death coding. Cause-of-death misclassification resulted in substantial underestimation of the true cumulative incidence of lung cancer deaths in NSCLC, mainly due to moderate levels of under-reporting of lung cancer deaths. Future studies that report the cumulative incidence of lung cancer deaths should correct for potential cause-of-death misclassification in the estimation, particularly among the elderly subpopulation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial/nonfinancial support: This work was supported by NIH Cancer Center Support Grant P30 CA008748. The sponsor played no role in any aspect of the study or manuscript.

Abbreviations:

AJCC	American Joint Committee on Cancer
CID	cumulative incidence of death
LC-CID	cumulative incidence of lung cancer deaths
Non-LC-CID	cumulative incidence of non-lung cancer deaths
NSCLC	non-small cell lung cancer
SEER	Surveillance, Epidemiology, and End Results

References

[1]. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, & Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer 2014; 91(7):1229–1235.

[2]. Tan KS, Eguchi T, Adusumilli PS. Competing risks and cancer-specific mortality: why it matters. Oncotarget 2018;9(7):7272. [PubMed: 29484108]

[3]. Mariotto AB, Noone AM, Howlader N, et al. Cancer survival: an overview of measures, uses, and interpretation. J Natl Cancer Inst Monogr 2014;2014(49):145–86. [PubMed: 25417231]

[4]. Perme MP, Estève J, Rachet B. Analysing population-based cancer survival – settling the controversies. BMC Cancer 2016;16(1):933. [PubMed: 27912732]

[5]. Howlader N, Mariotto AB, Woloshin S, Schwartz LM. Providing clinicians and patients with actual prognosis: cancer in the context of competing causes of death. J Natl Cancer Inst Monogr 2014 11 1;2014(49):255–64. [PubMed: 25417239]

[6]. Flanders WD. Inaccuracies of death certificate information. Epidemiology 1992;3:3–5. [PubMed: 1554807]

[7]. Hoel DG, Ron E, Carter R, Mabuchi K. Influence of death certificate errors on cancer mortality trends. J Natl Cancer Inst 1993;85(13):1063–1068. [PubMed: 8515493]

[8]. Doria-Rose VP, Marcus PM, Miller AB, Bergstralh EJ, Mandel JS, Tockman MS, Prorok PC. Does the source of death information affect cancer screening efficacy results? A study of the use of mortality review versus death certificates in four randomized trials. Clin Trials 2010;7(1):69–77. [PubMed: 20156958]

[9]. Gobbato F, Vecchiet F, Barbierato D, Melato M, Manconi R. Inaccuracy of death certificate diagnoses in malignancy: an analysis of 1,405 autopsied cases. Hum Pathol 1982;13(11):1036–1038. [PubMed: 7152507]

[10]. Lund JL, Harlan LC, Yabroff KR, Warren JL. Should cause of death from the death certificate be used to examine cancer-specific survival? A study of patients with distant stage disease. Cancer Invest 2010;28(7):758–764. [PubMed: 20504221]

[11]. Penson DF, Albertsen PC, Nelson PS, Barry M, Stanford JL. Determining cause of death in prostate cancer: are death certificates valid? J Natl Cancer Inst 2001;93(23):1822–1823. [PubMed: 11734600]

[12]. Percy C, Stanek E 3rd, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 1981;71(3):242–250. [PubMed: 7468855]

[13]. Rink M, Fajkovic H, Cha EK, Gupta A, Karakiewicz PI, Chun FK, Lotan Y, Shariat SF. Death certificates are valid for the determination of cause of death in patients with upper and lower tract urothelial carcinoma. Eur Urol 2012;61(4):854–855. [PubMed: 22226583]

[14]. Weinstock MA, Bogaars HA, Ashley M, Litle V, Bilodeau E, Kimmel S. Inaccuracies in certification of nonmelanoma skin cancer deaths. Am J Public Health 1992;82(2):278–281. [PubMed: 1739165]

[15]. Polednak AP. Inaccuracies in oral cavity–pharynx cancer coded as the underlying cause of death on US death certificates, and trends in mortality rates (1999–2010). Oral Oncol 2014;50(8):732–739. [PubMed: 24862544]

[16]. Doria-Rose VP, Marcus PM. Death certificates provide an adequate source of cause of death information when evaluating lung cancer mortality: an example from the Mayo Lung Project. Lung Cancer 2009;63(2):295–300. [PubMed: 18585822]

[17]. Edwards JK, Cole SR, Chu H, Olshan AF, Richardson DB. Accounting for outcome misclassification in estimates of the effect of occupational asbestos exposure on lung cancer death. Am J Epidemiol 2013;179(5):641–647. [PubMed: 24352593]

[18]. Hinchliffe SR, Abrams KR, Lambert PC. The impact of under and over-recording of cancer on death certificates in a competing risks analysis: a simulation study. Cancer Epidemiol 2013;37(1):11–19. [PubMed: 22999870]

[19]. Bakoyannis G, and Yiannoutsos CT. Impact of and correction for outcome misclassification in cumulative incidence estimation. PloS One 2015: e0137454.[dataset]

[20]. Surveillance, Epidemiology, and End Results (SEER) Program Populations (1973– 2014) Bethesda, MD: National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch; 2017.

[21]. Boer R, Ries L, van Ballegooijen M, Feuer E, Legler J, Habbema D. Ambiguities in Calculating Cancer Patient Survival: The SEER Experience for Colorectal and Prostate Cancer Bethesda, MD: Statistical Research and Applications Branch, National Cancer Institute; 2003 Technical Report No. 2003–05. http://srab.cancer.gov/reports. Accessed April 30, 2019.

[22]. Fritz A, Percy C, Jack A, et al., eds. International Classification of Diseases for Oncology 3rd ed. Geneva, Switzerland: World Health Organization; 2000.

[23]. Berlin NI, Buncher CR, Fontana RS et al. The National Cancer Institute Cooperative Early Lung Cancer Detection Program. Results of the initial screen (prevalence). Early lung cancer detection: introduction. Am Rev Respir Dis 1984; 130: 545–549. [PubMed: 6548343]

[24]. Fontana RS, Sanderson DR, Taylor WF, et al. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Mayo Clinic study. Am Rev Respir Dis 1984;130(4):561–565. [PubMed: 6091507]

[25]. Fontana RS, Sanderson DR, Woolner LB et al. Screening for lung cancer. A critique of the Mayo Lung Project. Cancer 1991; 67: 1155–1164. [PubMed: 1991274]

[26]. Frost JK, Ball WC Jr, Levin ML, et al. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. Am Rev Respir Dis 1984;130(4):549–554. [PubMed: 6091505]

[27]. Selikoff IJ, Seidman H. Use of death certificates in epidemiological studies, including occupational hazards: Variations in discordance of different asbestos-associated diseases on best evidence ascertainment. Am J Ind Med 1992;22(4):481–492. [PubMed: 1442783]

[28]. Hinchliffe SR, Rutherford MJ, Crowther MJ, et al. Should relative survival be used with lung cancer data? Br J Cancer 2012;106(11):1854–1859. [PubMed: 22555396]

[29]. Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. Int J Epidemiol 2010;39(2):598–610. [PubMed: 20142331]

[30]. Alpérovitch A, Bertrand M, Jougla E, et al. Do we really know the cause of death of the very old? Comparison between official mortality statistics and cohort study classification. Eur J Epidemiol 2009;24(11):669–675. [PubMed: 19728117]

[31]. Gamel JW, Vogel RL. Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients. Stat Methods in Med Res 2001;10(5):339–352. [PubMed: 11697226]

[32]. Bakoyannis G, Siannis F, Touloumi G. Modelling competing risks data with missing cause of failure. Stat Med 2010;29(30):3172–3185. [PubMed: 21170911]

[33]. Van Rompaye B, Goetghebeur E, Jaffar S. Design and testing for clinical trials faced with misclassified causes of death. Biostatistics 2010;11(3):546–558. [PubMed: 20212319]

[34]. Van Rompaye B, Jaffar S, Goetghebeur E. Estimation with Cox models: cause-specific survival analysis with misclassified cause of failure. Epidemiology 2012;23(2):194–202. [PubMed: 22317803]

[35]. Ha J, Tsodikov A. Isotonic estimation of survival under a misattribution of cause of death. Lifetime Data Anal 2012;18(1):58–79. [PubMed: 22094534]

[36]. Ebrahimi N The effects of misclassification of the actual cause of death in competing risks analysis. Stat in Med 1996;15(14): 1557–1566. [PubMed: 8855481]

Highlights:

- Causes of death recorded on death certificates are susceptible to errors.
- Lung cancer deaths can be misclassified as non-lung cancer deaths and vice-versa.
- Estimation of cumulative incidence of cancer-death requires accurate cause-of-death coding.
- Cause-of-death misclassification led to underestimation in cumulative incidence of lung cancer deaths.
- Bias in estimation increased with age, especially among those older than 75 years.

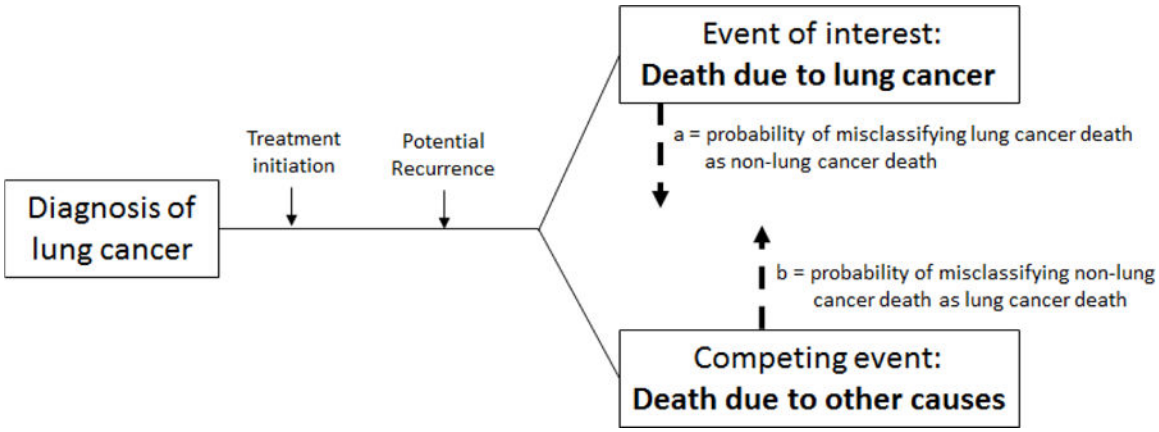

Figure 1. Diagram depicting the competing events for patients diagnosed with lung cancer.

In this context, deaths due to lung cancer are considered the events of interest and non-lung cancer deaths are considered competing events. Misclassification can occur when deaths due to lung cancer are misclassified as non-lung cancer deaths, leading to under-reporting of lung cancer deaths (*a*), or when non-lung cancer deaths are misclassified as lung cancer deaths, leading to over-reporting of lung cancer deaths (*b*).

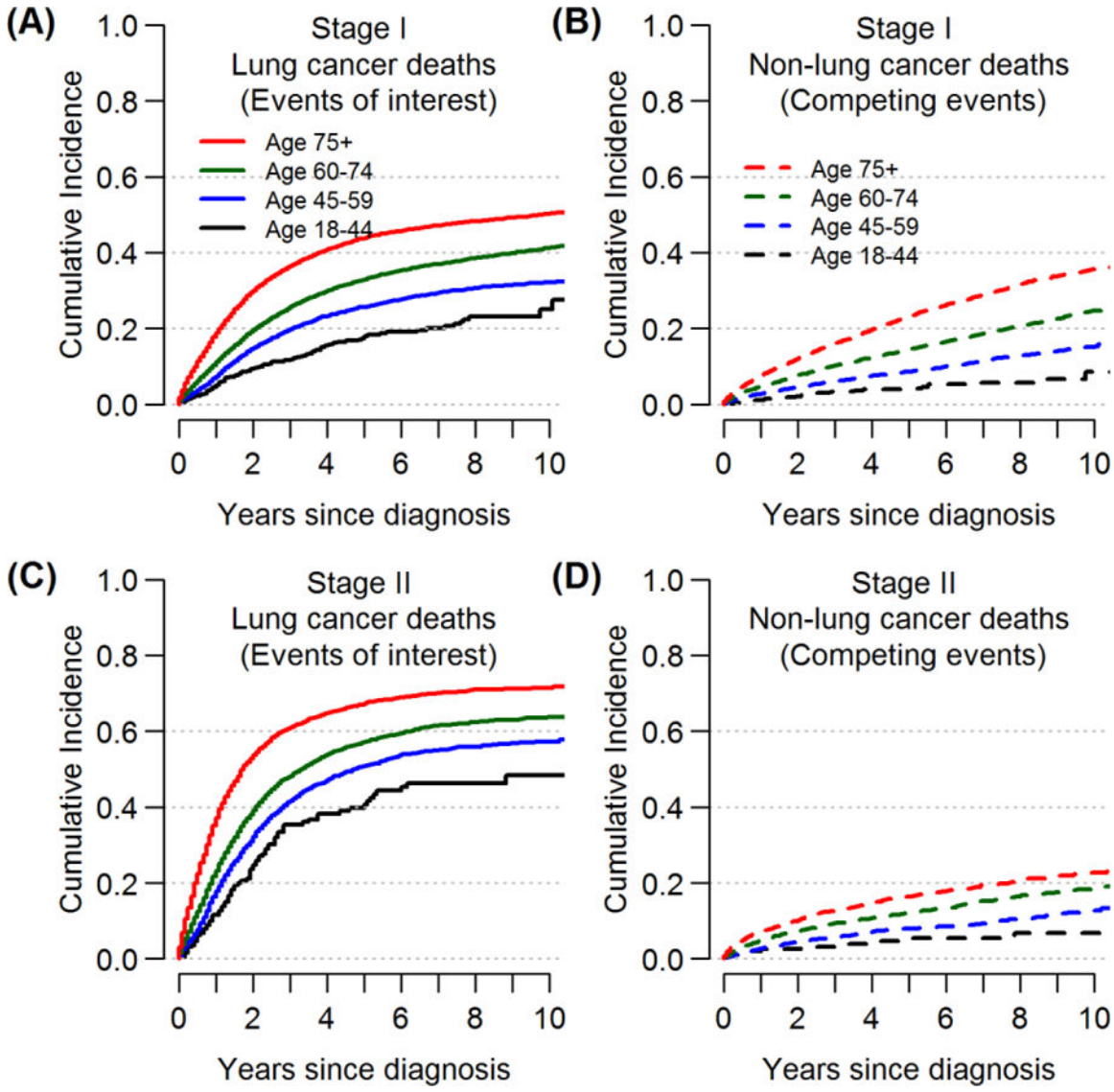

Figure 2.

Cumulative incidences of lung cancer deaths (the events of interest) and non-lung cancer deaths (the competing events) up to 10-years after diagnosis, by age and stage of disease.

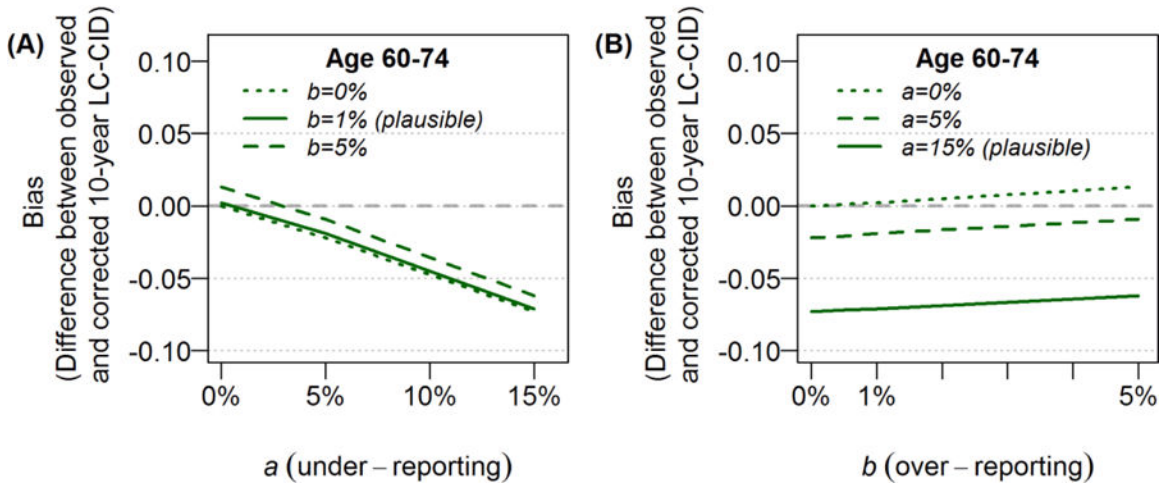

Figure 3. The impact cause-of-death misclassification on the 10-year cumulative incidence of lung cancer deaths (LC-CID) among patients aged 60 to 74 with stage I disease.

In each scenario, *a* refers to the probability of under-reporting lung cancer deaths, and *b* refers to the probability of over-reporting lung cancer deaths. Results are presented for increments in *b* given increasing values of *a* (left panel) or increments in *a* given increasing values of *b* (right panel). Based on literature, the plausible values of under-reporting and over-reporting are $a = 15\%$ and $b = 1\%$. The differences between observed 10-year LC-CIDs and corrected 10-year LC-CIDs were reported as biases in each scenario. A positive value of bias indicates that the observed 10-year LC-CID overestimated the corrected 10-year LC-CID; a negative value of bias indicates that the observed 10-year LC-CID underestimated the corrected 10-year LC-CID. The gray dashed line at y = 0 represents the ideal situation of no bias.

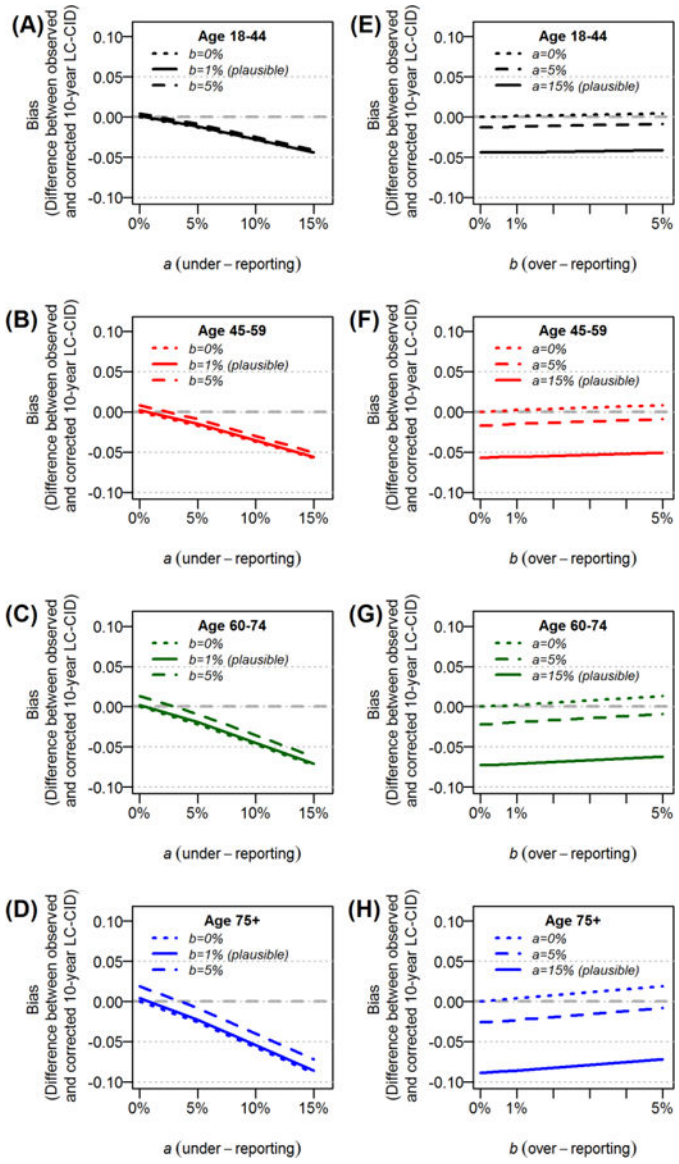

Figure 4.

The impact cause-of-death misclassification on the 10-year cumulative incidence of lung cancer deaths (LC-CID) among patients with stage I disease and increasing age at diagnosis.

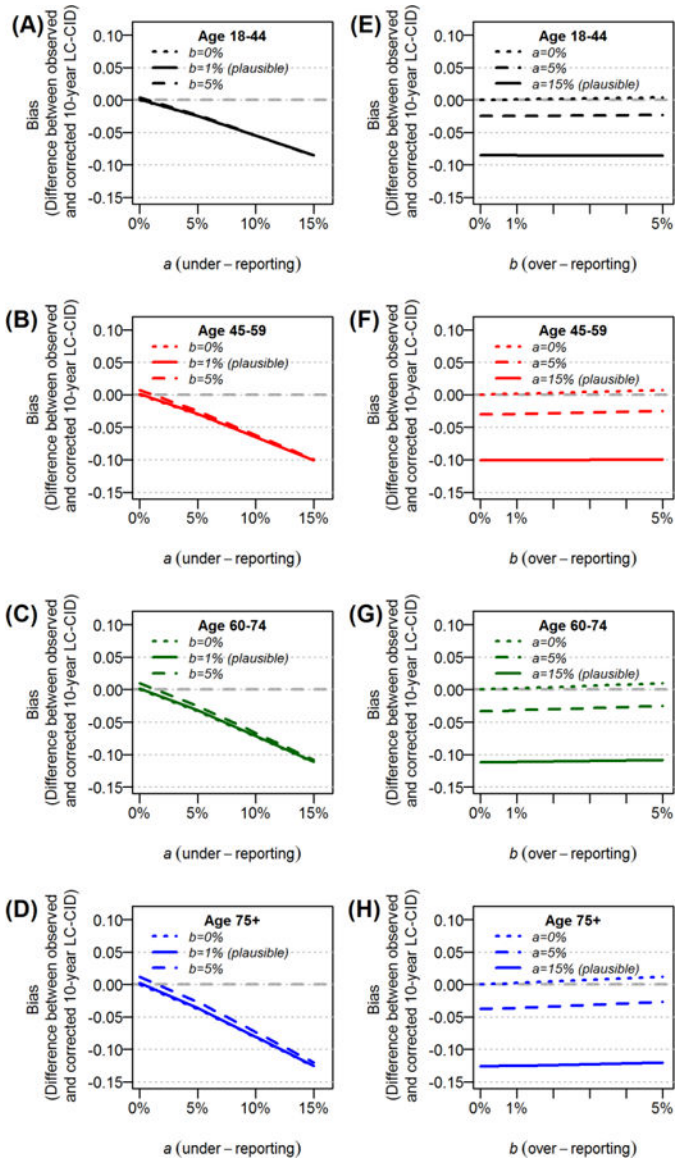

Figure 5.

The impact cause-of-death misclassification on the 10-year cumulative incidence of lung cancer deaths (LC-CID) among patients with stage II disease and increasing age at diagnosis.

Table 1:

Summary of clinical data from Surveillance, Epidemiology, and End Results (SEER) registry among patients with stage I or II non-small cell lung cancer

					Cause of death†			
					Lung cancer		Non-lung cancer	
Stage	Age**	% Female	Total No.	No. of All-cause deaths	Event No. (%)*	10-year LC-CID	Event No. (%)*	10-year non-LC-CID
I	18–44	61%	575	122	96 (79%)	25%	26 (21%)	9%
I	45–59	54%	8266	2548	1838 (72%)	32%	710 (28%)	15%
I	60–74	50%	25942	10813	7268 (67%)	41%	3545 (33%)	25%
I	75+	53%	18351	10850	6943 (64%)	50%	3907 (36%)	36%
II	18–44	50%	198	86	76 (88%)	48%	10 (12%)	7%
II	45–59	44%	2740	1448	1231 (85%)	57%	217 (15%)	13%
II	60–74	42%	6410	3942	3187 (81%)	63%	755 (19%)	18%
II	75+	47%	3697	2819	2245 (80%)	71%	574 (20%)	23%
Total		50%	66179	32628	22884 (70%)	47%	9744 (30%)	25%

† Based on the “SEER cause-specific death classification” variable in the SEER registry, subject to cause-of-death misclassification.

** Age at diagnosis.

* % among all-cause death within each subgroup ignoring censoring.

LC-CID, cumulative incidence of lung cancer deaths; non-LC-CID, cumulative incidence of non-lung cancer deaths.