Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# A Validated Register-Based Algorithm to Identify Patients Diagnosed with Recurrence of Surgically Treated Stage I Lung Cancer in Denmark

Linda Aagaard Rasmussen[1], Niels Lyhne Christensen[2], Anne Winther-Larsen[3], Susanne Oksbjerg Dalton[4,5], Line Flytkjær Virgilsen[1], Henry Jensen[1], Peter Vedsted[1]

[1]Research Unit for General Practice, Aarhus, Denmark; [2]Department of Pulmonary Medicine and Allergy, Aarhus University Hospital, Aarhus, Denmark; [3]Department of Clinical Biochemistry, Aarhus University Hospital, Aarhus, Denmark; [4]Survivorship and Inequality in Cancer, Danish Cancer Society Research Center, Copenhagen, Denmark; [5]Department of Clinical Oncology & Palliative Care, Zealand University Hospital, Næstved, Denmark

Correspondence: Linda Aagaard Rasmussen, Research Unit for General Practice, Bartholins Allé 2, Aarhus C, 8000, Denmark, Tel +45 8716 8365, Email linda.rasmussen@ph.au.dk



**Introduction:** Recurrence of cancer is not routinely registered in Danish national health registers. This study aimed to develop and validate a register-based algorithm to identify patients diagnosed with recurrent lung cancer and to estimate the accuracy of the identified diagnosis date.

**Material and Methods:** Patients with early-stage lung cancer treated with surgery were included in the study. Recurrence indicators were diagnosis and procedure codes recorded in the Danish National Patient Register and pathology results recorded in the Danish National Pathology Register. Information from CT scans and medical records served as the gold standard to assess the accuracy of the algorithm.

**Results:** The final population consisted of 217 patients; 72 (33%) had recurrence according to the gold standard. The median follow-up time since primary lung cancer diagnosis was 29 months (interquartile interval: 18–46). The algorithm for identifying a recurrence reached a sensitivity of 83.3% (95% CI: 72.7–91.1), a specificity of 93.8% (95% CI: 88.5–97.1), and a positive predictive value of 87.0% (95% CI: 76.7–93.9). The algorithm identified 70% of the recurrences within 60 days of the recurrence date registered by the gold standard method. The positive predictive value of the algorithm decreased to 70% when the algorithm was simulated in a population with a recurrence rate of 15%.

**Conclusion:** The proposed algorithm demonstrated good performance in a population with 33% recurrences over a median of 29 months. It can be used to identify patients diagnosed with recurrent lung cancer, and it may be a valuable tool for future research in this field. However, a lower positive predictive value is seen when applying the algorithm in populations with low recurrence rates.

**Keywords:** lung neoplasms, recurrence, algorithms, validation study, registries, Denmark

## Background

The number of cancer survivors has increased steadily during the past decade due to advances in diagnostic technologies and cancer treatments.[1,2] In Denmark, the incidence of lung cancer was 70 per 100,000 in men and 65 per 100,000 in women in 2020.[2] Surgery is the main curative treatment regimen,[3] and 35% of patients with lung cancer underwent surgery with curative intent in Denmark in 2020.[4] The reported recurrence rates after lung cancer surgery vary from 14% to 50%, depending on cancer stage and follow-up time.[5–8]

Surveillance for cancer recurrence is essential in cancer follow-up, and insight into the patient pathway in the period before cancer recurrence is important when caring for cancer survivors.[1,9] Nevertheless, the research is sparse in this

Received: 14 November 2022
Accepted: 15 February 2023
Published: 1 March 2023

field. Information on cancer recurrence is rarely available outside clinical trials. Even though electronic health records and tumor registries contain vast amounts of data, cancer recurrence is rarely captured routinely or consistently.[9] This makes it difficult to conduct population-level research in cancer recurrence. Consequently, during the past decade, various studies from different healthcare systems have developed register-based algorithms to identify patients diagnosed with recurrent cancer. Algorithms to identify women diagnosed with recurrent breast cancer are well described in the literature.[10] However, only a few studies have developed and validated such methods to identify patients diagnosed with recurrent lung cancer.[6,7,11]

Previous studies have demonstrated high validity of Danish national health registers to identify patients diagnosed with recurrent malignant melanoma, colorectal, breast, bladder and endometrial cancer, with sensitivities ranging from 85% to 100% and positive predictive values ranging from 72% to 95%.[12–18]

The aim of this study was to develop and validate a register-based algorithm to identify patients diagnosed with recurrent lung cancer in Denmark and to estimate the accuracy of the identified diagnosis date.

## Materials and Methods

We conducted a cohort study based on Danish national registers. Individual-level data were linked through the unique identification number assigned to all Danish citizens at birth or immigration.[19]

### Data Sources

Data were retrieved from four Danish national registers. The Danish Civil Registration System[19] provided data on vital status and migration. The Danish Cancer Register[20] provided information on cancer diagnoses, diagnosis date, and tumor stage. The Danish National Patient Register[21] provided information on tumor stage, procedure codes, and diagnosis codes for all cancer-related hospital contacts. The Danish National Pathology Register[22] provided data on Systematized Nomenclature of Medicine (SNOMED) classification registrations,[23] which allowed identification of malignant morphology (codes M8* and M9*). The fifth digit of the morphology code indicates behavior (eg, 4: "direct spread to surrounding tissue", 6: "malignant metastasis", and 7: "malignant recurrence").

### Gold Standard

We included patients identified in the Danish Lung Cancer Register with stage I lung cancer during 2011–2014 who had received lung cancer surgery with curative intent. This population originated from a study by Christensen et al, which also included patients treated by stereotactic body radiotherapy and other primary cancer treatment regimes.[24] The patients were staged according to the Danish lung cancer guidelines, which are in line with international recommendations.[25] The majority of the patients with recurrence had a positron emission tomography with 2-deoxy-2-[fluorine-18] fluoro-D-glucose (FDG PET) performed. Additional diagnostic tests performed on suspicion of recurrence were mainly focused on the anatomic site under suspicion and included cerebral magnetic resonance imaging (MRI) or computed tomography (CT), biopsies, endobronchial ultrasound (EBUS) and endoscopic ultrasound (EUS). Patients diagnosed with recurrent lung cancer were identified through a medical record review (performed in May 2016) including imaging results and/or clinical evaluation. Distinction between recurrence and a second primary lung cancer was based on the conclusion from the multidisciplinary team conference that established the diagnostic work-up for the recurrence. Recurrence dates were registered as the number of months from the diagnosis date of the primary lung cancer to the month when the diagnostic procedures for recurrence had been undertaken.

### Study Population

Patients from the gold standard population were eligible for inclusion if registered in the Danish Cancer Register with a lung cancer diagnosis (C34* in the International Classification of Diseases, 10th revision (ICD-10)) and if aged 18 years or older at the lung cancer diagnosis date.

Patients were excluded in case of register-based indication of a distant metastasis within 90 days of the initial lung cancer diagnosis date identified in the Danish Cancer Register or within 90 days of the curative treatment initiation date identified in the Danish National Patient Register. Indications of a distant metastasis were selected diagnosis codes (ICD-10: C76*-C79* and C34xM) or SNOMED codes (M8*-M9* and 6 in the 5th digit), or a distant tumor stage based on the Union for International Cancer Control (UICC) classification of malignant tumors.[26] Further, patients were excluded if we were unable to identify procedure codes indicating treatment with curative intent in the Danish National Patient Register. Curative intent surgery was defined as procedure codes for wedge resection (KGDB1*), segmental resection (KGDB2*), other lung resection (KGDB96 or KGDB97), lobectomy (KGDC*), or pneumonectomy (KGDD*) combined with a diagnosis code of lung cancer (ICD-10: C34* except C349X; lung cancer recurrence) and registered less than 90 days after the diagnosis date. Adjuvant chemotherapy was defined as procedure codes for chemotherapy (BWHA*) combined with a diagnosis code of lung cancer (ICD-10: C34* except C349X; lung cancer recurrence), intrathoracic lymph nodes (ICD-10: C771) or lymph nodes, unspecified (ICD-10: C779), with a maximum of one registration per day, a maximum of 30 days between registrations, a maximum of nine procedure codes in total, and a maximum of 60 days from the date of surgery to the first date of chemotherapy.

## Algorithm

The algorithm was constructed similarly to the previously developed algorithms to identify cancer recurrence from malignant melanoma, bladder, breast, and endometrial cancer by Rasmussen et al[13–16] (Figure 1). The end date of primary lung cancer treatment was defined as the date of lung cancer surgery or the date of the last chemotherapy procedure code in case of adjuvant chemotherapy. A subsequent period with no register-based evidence of ongoing malignant disease was required to prevent inclusion of patients with residual disease after completed lung cancer treatment. The final day of this period was 90 days after surgery or 30 days after ended adjuvant chemotherapy treatment, whichever came last. Indicators of ongoing disease was 1) malignant morphology (SNOMED codes M8* and M9*), 2) new diagnosis codes indicating malignant disease (ICD-10: C00*- C96* and D37*-D48*excluding C44* (non-melanoma skin cancer) and C34* (lung cancer)), 3) procedure codes for radiotherapy (BWGC*) or chemotherapy (BWHA*) combined with a malignant diagnosis code (ICD-10: C00*-C96* and D37*-D48*), and 4) UICC stage IV.

After the period with no register-based evidence of ongoing malignant disease, the algorithm searched for indicators of cancer recurrence (Figure 1). A patient was defined as being diagnosed with lung cancer recurrence if one of the following six indicators was present:

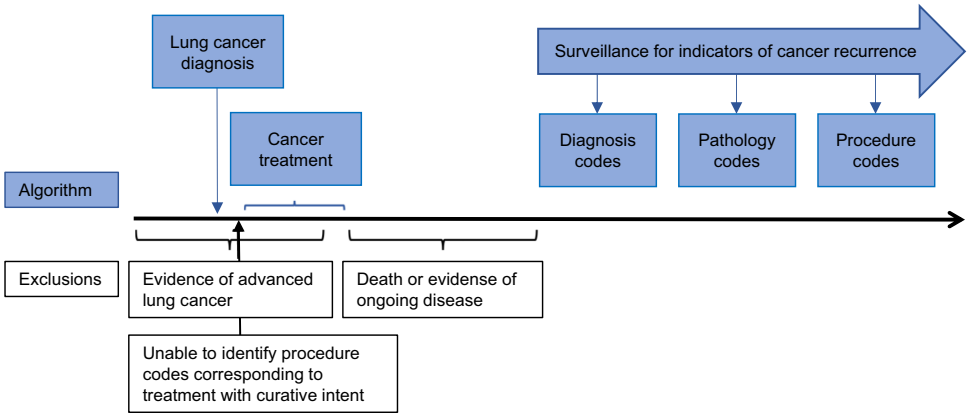


**Figure 1** Schematic overview of the algorithm.

1. ICD-10: C349X (lung cancer recurrence diagnosis);
2. ICD-10: C76*-C79* or C34xM (metastasis diagnosis) and no new primary cancer registered after the conclusion of primary lung cancer treatment;
3. SNOMED morphology codes M8*-M9* and 7 (malignant recurrence) in the fifth digit;
4. SNOMED morphology codes M8*-M9* and 4 (direct spread to surrounding tissue) or 6 (malignant metastasis) in the fifth digit and a morphology similar to a morphology code registered within 90 days of the primary lung cancer diagnosis date or date of lung cancer surgery;
5. Radiotherapy or chemotherapy procedure codes combined with a diagnosis code indicating lung cancer (ICD-10: C34*);
6. Radiotherapy or chemotherapy procedure codes combined with a diagnosis code indicating metastases (ICD-10: C76*-C79* or C34xM) and no new primary cancer registered after the conclusion of primary lung cancer treatment.

Indicators of recurrence were disregarded if they appeared after ended follow-up in the gold standard. The recurrence date estimated by the algorithm was defined as the first date with a registration of an indicator of recurrence.

## Statistical Analyses

The concordant and discordant frequencies between recurrences identified by the algorithm and by the gold standard were used to compute the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with 95% confidence interval (CI). The agreement between the date of recurrence identified by the algorithm and the date identified by the gold standard was measured by Lin's concordance correlation coefficient (CCC) score.[27] The agreement is considered "poor" for CCC < 0.90, "moderate" for CCC = 0.90–0.95, "substantial" for CCC > 0.95, and "almost perfect" for CCC > 0.99.[27] Furthermore, we analyzed the proportion of recurrence dates estimated by the algorithm to be on the same date as the gold standard recurrence date, and within 7, 30, and 60 days of the gold standard recurrence date. We simulated the PPV and the NPV of the algorithm's sensitivity and specificity in populations with a recurrence rate of 15%, 25%, and 50%, respectively.

Finally, we conducted two sub-analyses. First, we excluded patients with a cancer diagnosis recorded in the Danish Cancer Register prior to the lung cancer diagnosis to investigate if the algorithm performed differently in a population with no history of cancer. Second, the performance of the algorithm was estimated with second primary lung cancers and recurrences pooled as the outcome, to investigate the performance of the algorithm for studies, where the type of the new cancer event is not the primary focus, eg, studies on diagnostic trajectories in general practice.

We intended to develop and validate an algorithm to identify patients diagnosed with recurrence of non-metastatic small-cell lung cancer (SCLC). However, after exclusions, our study population comprised very few patients, and we decided not to proceed with this algorithm.[28]

# Results

The final study population comprised 217 patients, hereof 72 (33%) with recurrence according to the gold standard (Figure 2 and Table 1). For primary lung cancer surgery, 75% of the study population had a lobectomy alone, 5% had a bi-lobectomy or a lobectomy and a sleeve or segment resection, 18% had a sleeve or segment resection, and less than 3% had a pneumonectomy.

The median follow-up time since the primary lung cancer diagnosis date was 29 months (interquartile interval: 18–46). The algorithm identified 60 of the 72 recurrences according to the gold standard and an additional 9 false positives, hereof 5 with a second primary lung cancer according to the gold standard (Table 2). The algorithm reached a sensitivity of 83.3% (95% CI: 72.7–91.1) and a specificity of 93.8% (95% CI: 88.5–97.1) (Table 3). The agreement between the recurrence dates generated by the algorithm and the dates generated by the gold standard achieved a CCC score of 0.836 (95% CI 0.823–0.943), and 70% of the recurrence dates estimated by the algorithm were found to be within 60 days of the gold standard recurrence date (Table 4). The PPV span from 70.3% to 93.1% when simulated in populations with recurrence rates of 15% to 50% (Table 5).

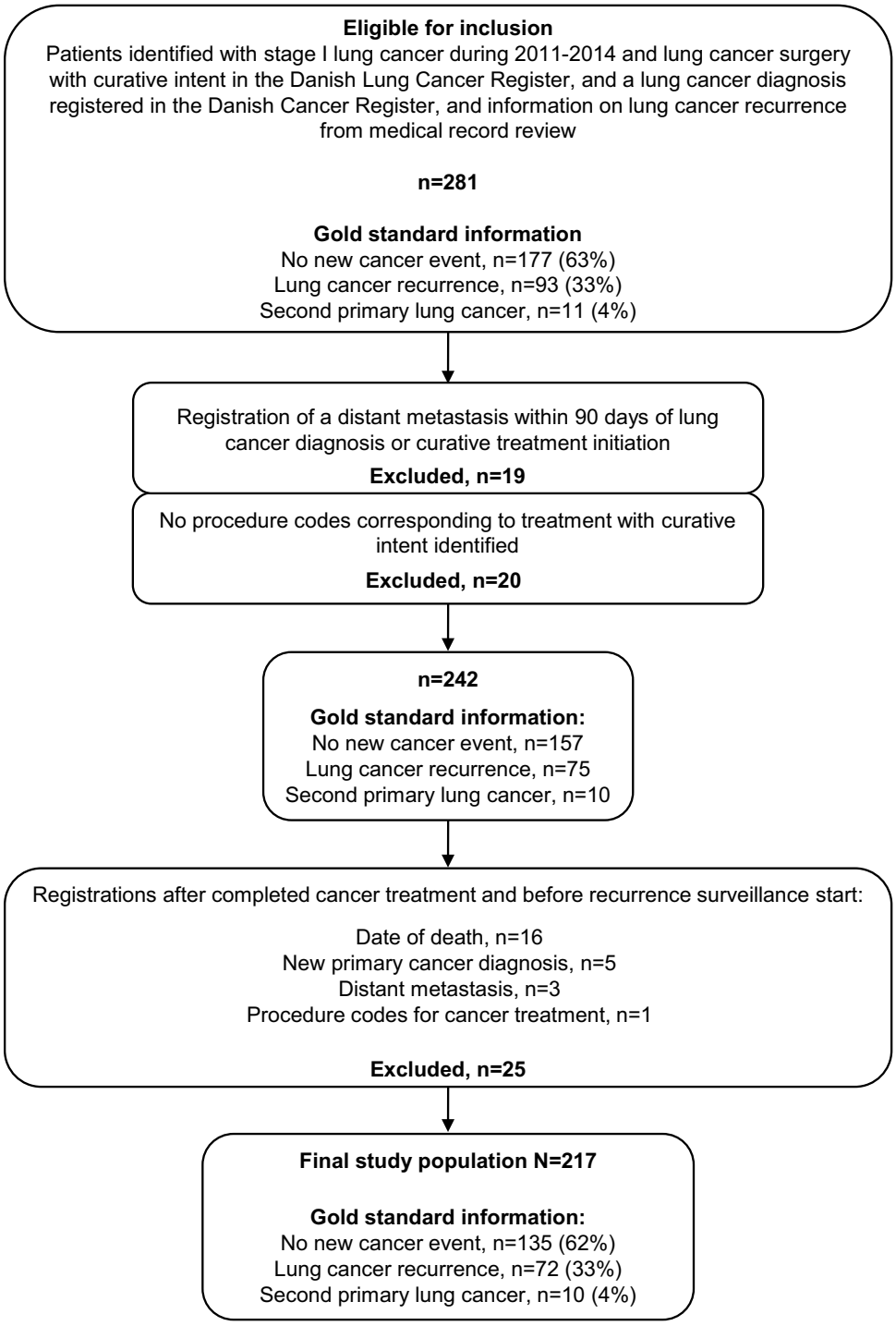


**Figure 2** Flowchart of study population.

In the first sub-analysis, the algorithm performed similarly to its performance in the main analysis when restricted to patients with no history of cancer prior to lung cancer (Table 3). In the second sub-analysis, the sensitivity decreased to 79.3% (95% CI: 68.9–87.4) and the specificity increased to 97.0% (95% CI: 92.5–99.2) when we pooled second primary lung cancer and cancer recurrence in the gold standard (Table 3).

**Table 1** Characteristics of the 217 Included Patients with Stage I Lung Cancer, Stratified on Cancer Recurrence Status in the Gold Standard[a]

| Population Characteristics | Cancer Recurrence | No Cancer Recurrence |
|---|---|---|
| | n (%) | n (%) |
| Number | 72 (100) | 145 (100) |
| Sex | | |
| Female | 27 (38) | 58 (40) |
| Male | 45 (63) | 87 (60) |
| Age, median (IQI[b]) | 74 (68;78) | 73 (67;76) |
| Comorbidity[c] | | |
| Low | 29 (40) | 60 (41) |
| Moderate | 23 (32) | 60 (41) |
| Severe | 20 (28) | 25 (17) |
| Tumor histology | | |
| Adenocarcinoma | 79 (54) | 33 (46) |
| Squamous cell carcinoma | 37 (26) | 25 (35) |
| Non-small cell carcinoma | 24 (17) | 8 (11) |
| Other | 5(4) | 6 (8) |
| Follow-up time[d], months (IQI) | 18 (11;27) | 36 (23;59) |

**Notes**: [a]Numbers are stated as n (%) if nothing else is stated; [b]Interquartile interval; [c]Charlson Comorbidity Index score calculated at primary lung cancer diagnosis date, divided into score 0 (low), scores 1–2 (moderate) and scores 3+ (severe); [d]Time from primary lung cancer diagnosis to the first of the following events: recurrence, emigration, death, or end of study (1 May 2016).

**Table 2** Concordance of Recurrence of Lung Cancer Identified by the Gold Standard and the Algorithm

| Status by Algorithm | Status by Gold Standard | | | |
|---|---|---|---|---|
| | Cancer Recurrence | No Cancer Recurrence | | Total |
| | | No New Cancer Event | New Primary Lung Cancer | |
| Cancer recurrence | 60 | 4 | 5 | 69 |
| No cancer recurrence | 12 | 131 | 5 | 148 |
| Total | 72 | 135 | 10 | 217 |

**Table 3** Performance of the Algorithm to Identify Recurrence from Lung Cancer in Patients Treated for Stage I Lung Cancer with Curative Intent

| Algorithm Performance | Main Analysis n=217 | Excluding 31 Patients with a History of Cancer Prior to Lung Cancer, n=186 | Second Primary Lung Cancer and Cancer Recurrence Pooled in the Gold Standard, n=217 |
|---|---|---|---|
| | % (95% CI) | % (95% CI) | % (95% CI) |
| Sensitivity | 83.3 (72.7–91.1) | 85.0 (73.4–92.9) | 79.3 (68.9–87.4) |
| Specificity | 93.8 (88.5–97.1) | 93.7 (87.9–97.2) | 97.0 (92.6–99.2) |
| Positive predictive value | 87.0 (76.7–93.9) | 86.4 (75.0–94.0) | 94.2 (85.8–98.4) |
| Negative predictive value | 91.9 (86.3–95.7) | 92.9 (87.0–96.7) | 88.5 (82.2–93.2) |

**Abbreviation**: CI, confidence interval.

**Table 4** Concordance of Cancer Recurrence Date in the Gold Standard and Estimated by the Algorithm

| Estimated Date by the Algorithm | % (95% CI) |
|---|---|
| On same date | 0 |
| Within 7 days | 8.3 (3.6–18.9) |
| Within 14 days | 20.0 (11.9–32.5) |
| Within 30 days | 40.0 (29.9–53.4) |
| Within 60 days | 70.0 (58.3–81.0) |

**Abbreviation**: CI, confidence interval.

**Table 5** Simulated Positive Predictive and Negative Predictive Value of the Sensitivity and Specificity of the Algorithm to Identify Recurrence from Lung Cancer in Constructed Populations with Different Recurrence Rates

| Algorithm Performance | Population n=10,000 Recurrence Rate 15% | Population n=10,000 Recurrence Rate 25% | Population n=10,000 Recurrence Rate 50% |
|---|---|---|---|
| | % (95% CI) | % (95% CI) | % (95% CI) |
| Positive predictive value | 70.3 (68.2–72.5) | 81.8 (80.2–83.2) | 93.1 (92.3–93.8) |
| Negative predictive value | 96.9 (96.6–97.3) | 94.4 (93.8–94.9) | 84.9 (83.9–85.8) |

**Abbreviation**: CI, confidence interval.

# Discussion

## Main Findings

Based on national registers, we developed and validated an algorithm to identify patients diagnosed with recurrence from surgically treated stage I lung cancer. The algorithm reached a sensitivity of 83.3 (95% CI: 72.7–91.1) and a PPV of 87.0 (95% CI: 76.7–93.9), and the algorithm estimated 70% of recurrence dates within 60 days of the gold standard recurrence date.

# Strengths and Limitations

The most important strength of the study was the linkage of individual-level data in high-quality population-based registers combined with the free and equal access to healthcare for all inhabitants in Denmark. This increased the applicability of the algorithm in future studies, and it limited the risk of both information bias and selection bias.[29] However, misclassification of recurrence and recurrence dates from missing or incorrect registrations did occur. Lung cancer patients tend to be a fragile population with a higher prevalence of comorbidities than seen for other cancers,[3] which might contraindicate biopsy and cancer treatment in case of recurrent disease. This increased the risk of being missed by the algorithm. Further, false positives caused a PPV of 87%. However, when we included second primary lung cancers with cancer recurrence in the gold standard, the PPV increased to 94%.

Another important strength was that the gold standard population was derived from a population-based database, which increased the representativeness of the study population to the entire population of early-stage lung cancer patients treated surgically with curative intent in Denmark. Patients with recurrence from stage II and III cancers are registered with the same diagnosis, procedure, and pathology codes as patients with stage I cancer,[15] and previous studies of malignant melanoma,[16] bladder,[13] breast,[14] and colorectal cancers[12] have shown similar performance of the indicators of recurrence, also in patients with stage II and III cancer. Thus, we argue that the algorithm is applicable to the entire population of patients treated with curative-intent surgery for stage I–III lung cancer. However, validation in a population with stage II–III lung cancer the algorithm would be preferable, and it could reveal an increased PPV due to an expected higher recurrence rate. The inclusion of patients with a history of cancer prior to the lung cancer diagnosis might have introduced a risk of misclassifying registrations related to recurrence of the previous cancer as registrations related to recurrence of lung cancer. However, the algorithm performed similarly to the algorithm that excluded patients with prior

cancer. This broadens the applicability of the algorithm to cover also cases with a second primary cancer, which account for up to 16% of all cancer diagnoses.[30]

An important limitation was the unsuccessful attempt to develop an algorithm to identify recurrence of lung cancer treated with radiotherapy in curative doses. The primary reason for this was that more than 50% of the eligible population had register-based evidence of a distant metastasis within 90 days of the lung cancer diagnosis date or the date of curative treatment initiation, and additional 23% had register-based evidence of ongoing malignant disease after completed cancer treatment. Consequently, research in lung cancer recurrence remains challenging in patients treated with radiotherapy and is restricted to actively followed patients, eg, in clinical trials.

Another limitation of the study was the relatively low number of analysed patients which leaded to large CIs of the estimates. We were unable to identify procedure codes indicating curative treatment in the Danish National Patient Register for 20 of 281 eligible lung cancer patients undergoing curative intent surgery (7%). In 12 cases, we were unable to identify procedure codes for curative intent surgery combined with a diagnosis of C34*. In the remaining 8 cases, the identified code for curative surgery was registered more than 90 days after the lung cancer diagnosis date in the Danish Cancer Register. Thirteen (65%) of these cases had a lung cancer recurrence according to the gold standard, which was a considerably larger proportion compared to the recurrence rate in the final population (33%). This might hold a risk of underestimating the absolute number of cases with recurrence in future studies using the algorithm to sample patients with lung cancer recurrence. Further, it might produce bias if these excluded patients differ from the source population on parameters relevant for study outcomes, such as survival. Another 44 patients were excluded due to death or register-based indicators of malignant disease within 90 days after cancer surgery or within 30 days of ended adjuvant therapy. Nineteen of these were excluded due to a registration of a distant metastasis. This corresponds with findings from 2019 to 2020 based on the Danish Lung Cancer Register[4] in which six percent of surgically treated lung cancer patients had different preoperative clinical staging and postoperative pathological staging.

## Comparison with Other Studies

The algorithm was based on the same data sources and was constructed similarly to other algorithms developed and validated by our research group to identify recurrence from bladder, breast, and endometrial cancer and from malignant melanoma.[12,14–18] However, the sensitivity at 83% was lower than found in these previous studies with sensitivities ranging from 85% to 97%. This might be caused by limited treatment options or reduced rationale for biopsy in a fragile population with high comorbidity.[3] For example, only 34% of the lung cancer recurrences were identified based on a pathology code (data not shown) compared to 52% for malignant melanoma[16] and 71% for endometrial cancer.[15] Further, the 12 false negatives were older (mean 77 years (IQI: 69–80)) than the true positives (mean 73 years (IQI 68–77)), and more false negative than true positives had severe comorbidity (33% vs 27%) (data not shown). The algorithms by Hassett et al[6,7] to identify patients diagnosed with recurrence from lung cancer were based on data from medical claims and included indicators from diagnosis codes for secondary malignancy and procedure codes for chemotherapy. These algorithms reached sensitivities in the range 77–91% and specificities in the range 70%–95%. However, the PPV decreased to 74% in the versions aiming for a high sensitivity. We intended to develop an algorithm with both high sensitivity and high PPV. If the purpose of a study is to identify as many patients with recurrence as possible, regardless of the consequences for false positives and a low PPV, more indicators of recurrence may be added to the algorithm, eg, all malignant pathology test results, lung cancer diagnosis codes, and procedure codes for chemotherapy and radiotherapy regardless of the indication diagnosis. Contrary, if it is important with a high PPV, more restrictions could be posed to the indicators, eg, limiting the indicators to the morphology codes.

The estimated recurrence dates were less accurate for lung cancer than found in similar analyses made by our research group for cancer of the bladder, breast, endometrial, and melanoma.[13–16] The date of recurrence in the gold standard was estimated as the first day in the nearest month in which the diagnostic workup for recurrence was undertaken, ie, no exact date was estimated from a biopsy test or procedure code. This may have contributed to the poorer agreement found in this study, and the actual agreement may be better than reported. However, when studying event-free survival, it is recommended to identify the correct recurrence date from medical record review or interviews.

Similar algorithms published outside Denmark are primarily based on data from medical claims.[10] Thus, these are restricted to specific populations, geographic areas, or insurance groups. This limits their generalizability to entire populations, whereas algorithms based on the population-based nationwide Danish health registers allow for national applicability. The majority of register-based algorithms to identify patients diagnosed with cancer recurrence have focused on solid cancers treated by surgery.[6,7,10–17] We were unable to develop an algorithm to identify patients diagnosed with recurrence from SCLC after treatment by radiotherapy in curative doses. It is more complex to identify a curative treatment course and to distinguish between curative and palliative oncological treatment compared to surgical treatment, which is often one single surgery procedure code. This was also reported in another of our previous studies[13] on patients diagnosed with recurrence after bladder cancer, which included both patients treated by surgery and by radiotherapy. This algorithm showed substantially superior performance in patients undergoing surgical treatment for bladder cancer.

### Interpretation and Implications

The algorithm is a time-saving method for identifying patients diagnosed with lung cancer recurrence in large populations. The population-based design enables sampling of large populations, which allows for detailed analyses with high statistical precision. Compared to clinical trials, there are no exclusion criteria related to age and comorbidity, no underrepresented sub-groups due to active enrollment, and no loss to follow-up. These features increase the generalizability of our findings to the entire population of patients treated with curative-intent surgery for lung cancer.

The recurrence rate should be considered in studies using the algorithm to sample patients diagnosed with recurrence since the PPV decreased with decreasing recurrence rate (Table 5). It is especially important to consider the false-positive rate when planning studies with active contact to the patients, in which case one would be ethically obligated to avoid as many false positives as possible, and one should aim for a high specificity. However, in large-population epidemiological studies, misclassification of recurrence status will most often lead to an underestimated difference in the outcome between the explored and the unexplored, unless the misclassified patients differ substantially by the outcome of interest.[31] If the purpose of a study is to identify as many patients with recurrence as possible, more indicators of recurrence may be added to the algorithm to increase the sensitivity.

The proposed algorithm facilitates epidemiological research in the field of lung cancer recurrence. Research in disease-free survival after cancer treatment is an important supplement to investigations in mortality and survival.[9] Further, the algorithm offers an opportunity to identify the characteristics of the population with recurrence and enhance the evidence on the diagnostic pathway for lung cancer recurrence. Such new knowledge may inform the organization of cancer follow-up and improve the surveillance of recurrence.[32]

## Conclusion

We developed and validated a register-based algorithm to identify patients diagnosed with recurrence of early-stage lung cancer. The algorithm showed good performance; it identified 83% of all recurrent cancers, and 70% of the recurrence diagnosis dates were estimated within 60 days of the gold standard. Algorithms using population-level data may contribute with evidence on potential outcomes for patients with recurrence outside clinical trials. The algorithm may serve as a valuable resource for research in the field of lung cancer recurrence.

## Abbreviations

CCC, concordance correlation coefficient; CI, confidence interval; ICD-10, International Classification of Diseases, 10th revision; NPV, negative predictive value; PPV, positive predictive value; SCLC, small-cell lung cancer; SNOMED, Systematized Nomenclature of Medicine; UICC, Union for International Cancer Control.

## Data Sharing Statement

The data supporting the findings of this study is stored and maintained electronically on servers at Statistics Denmark. The data is only accessible through a secure virtual private network (VPN) and only for pre-approved collaborative

partners. The data is not publicly available due to the Danish data protection legislation as the data contains information that could compromise the privacy of the research participants.

## Ethical Approval

The project is approved and registered as "The patient pathway for cancer recurrence" (study 1, id 119) in the Record of Processing Activities at the Research Unit for General Practice in Aarhus in accordance with the provisions of the General Data Protection Regulation (GDPR). According to the Danish committee law, §14, 2, the study required no approval from the Committee on Health Research Ethics of the Central Denmark Region as "no biomedical intervention was performed".[33]

## Author Contributions

LAR, LFV, HJ, and PV conceived the concept of the study. AWL, NLC, and SOD provided data on the gold standards. LAR was responsible for the data acquisition, data management, statistical analyses and drafted the manuscript. All authors contributed with significant contribution to the reported work, including the analyses, interpretation of data and substantial and critical review of the article. All authors read and approved the final version of the manuscript and agreed on the submission to Clinical Epidemiology. All authors agree to take responsibility and be accountable for the contents of the article.

## Funding

The study was supported by the Research Foundation for General Practice in the Central Denmark Region. The funder was not involved in any aspects of the study.

## Disclosure

The authors declare that they have no conflicts of interest in this work.

## References

1. Rubin G, Berendsen A, Crawford SM, et al. The expanding role of primary care in cancer control. *Lancet Oncol*. 2015;16:1474–5488.
2. The Danish National Board of Health [Homepage on the internet]. Cancer incidence in Denmark 2019 (in Danish); 2021. Available from: https://sundhedsdatastyrelsen.dk/da/find-tal-og-analyser/tal-og-analyser/sygdomme-og-behandlinger/kraeft/kraeft_nyetilfaelde_aarsrapport. Accessed October 17, 2022.
3. Leduc C, Antoni D, Charloux A, Falcoz PE, Quoix E. Comorbidities in the management of patients with lung cancer. *Eur Respir J*. 2017;49(3):1601721.
4. The Danish Lung Cancer group, The Danish Lung Cancer Register [Homepage on the internet]. Danish lung cancer register - annual national report 2019–2020 (in Danish); 2021. Available from: https://www.lungecancer.dk/wp-content/uploads/2021/10/%C3%85rsrapport-20192020_netudgave.pdf. Accessed October 17, 2022.
5. Subotic D, Van Schil P, Grigoriu B. Optimising treatment for post-operative lung cancer recurrence. *Eur Respir J*. 2016;47(2):374–378.
6. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Med Care*. 2017;55(12):e88–e98.
7. Hassett MJ, Ritzwoller DP, Taback N, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large contemporary cohorts. *Med Care*. 2014;52(10):e65–e73.
8. Takenaka T, Tagawa T, Kohno M, et al. Consideration of the optimal surgical procedure based on the risk of recurrence in clinical stage 0 or IA lung adenocarcinoma. *Anticancer Res*. 2022;42(2):1137–1142.
9. Warren JL, Yabroff KR. Challenges and opportunities in measuring cancer recurrence in the United States. *J Natl Cancer Inst*. 2015;107(8):djv134–djv134.
10. Izci H, Tambuyzer T, Tuand K, et al. A systematic review of estimating breast cancer recurrence at the population level with administrative data. *J Natl Cancer Inst*. 2020;112(10):djaa050.
11. Uno H, Ritzwoller DP, Cronin AM, Carroll NM, Hornbrook MC, Hassett MJ. Determining the time of cancer recurrence using claims or electronic medical record data. *JCO Clin Cancer Inform*. 2018;2(2):1–10.
12. Lash TL, Riis AH, Ostenfeld EB, Erichsen R, Vyberg M, Thorlacius-Ussing O. A validated algorithm to ascertain colorectal cancer recurrence using registry resources in Denmark. *Int J Cancer*. 2015;136(9):2210–2215.
13. Rasmussen LA, Jensen H, Virgilsen LF, Jensen JB, Vedsted P. A validated algorithm to identify recurrence of bladder cancer: a register-based study in Denmark. *Clin Epidemiol*. 2018;10:1755–1763.
14. Rasmussen LA, Jensen H, Flytkjær Virgilsen L, Jellesmark Thorsen LB, Vrou Offersen B, Vedsted P. A validated algorithm for register-based identification of patients with recurrence of breast cancer—based on Danish Breast Cancer Group (DBCG) data. *Cancer Epidemiol*. 2019;59:129–134.

15. Rasmussen LA, Jensen H, Virgilsen LF, et al. Identification of endometrial cancer recurrence – a validated algorithm based on nationwide Danish registries. *Acta Oncol*. 2021;60(4):1–7.
16. Rasmussen LA, Jensen H, Virgilsen LF, Hölmich LR, Vedsted P. A validated register-based algorithm to identify patients diagnosed with recurrence of malignant melanoma in Denmark. *Clin Epidemiol*. 2021;13:207–214.
17. Pedersen RN, Öztürk B, Mellemkjær L, et al. Validation of an algorithm to ascertain late breast cancer recurrence using Danish medical registries. *Clin Epidemiol*. 2020;12:1083–1093.
18. Cronin-Fenton D, Kjærsgaard A, Nørgaard M, et al. Breast cancer recurrence, bone metastases, and visceral metastases in women with stage II and III breast cancer in Denmark. *Breast Cancer Res Treat*. 2018;167(2):517–528.
19. Pedersen CB. The Danish civil registration system. *Scand J Public Health*. 2011;39(7 Suppl):22–25.
20. Gjerstorff ML. The Danish cancer registry. *Scand J Public Health*. 2011;39(7_suppl):42–45.
21. Sandegaard JL, Schmidt SAJ, Sørensen HT, et al. The Danish national patient registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490.
22. Erichsen R, Lash TL, Hamilton-Dutoit SJ, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish national pathology registry and data bank. *Clin Epidemiol*. 2010;2:51–56.
23. SNOMED international [Homepage on the internet]. SNOMED international. Available from: https://www.snomed.org/. Accessed October 17, 2022.
24. Christensen NL, Dalton SO, Mellemgaard A, Christensen J, Kejs AMT, Rasmussen TR. Assessing the pattern of recurrence in Danish stage I lung cancer patients in relation to the follow-up program: are we failing to identify patients with cerebral recurrence? *Acta Oncol*. 2018;57(11):1556–1560.
25. Christensen NL, Jekunen A, Heinonen S, Dalton DO, Rasmussen TR. Lung cancer guidelines in Sweden, Denmark, Norway and Finland: a comparison. *Acta Oncol*. 2017;56(7):943–948.
26. Union for International Cancer Control [Homepage on the internet]. TNM classification of malignant tumours; 2018. Available from: https://www.uicc.org/resources/tnm. Accessed October 17, 2022.
27. McBride GB. *A Proposal for Strength of Agreement Criteria for Lin's Concordance Correlation Coefficient*. Hamilton, New Zealand: National Institute of Water & Atmospheric Reseach; 2005.
28. Winther-Larsen A, Hoffmann L, Moeller DS, Khalil AA, Knap MM. Evaluation of factors associated with loco-regional failure and survival in limited disease small cell lung cancer patients treated with chemoradiotherapy. *Acta Oncol*. 2015;54(9):1574–1581.
29. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol*. 2019;11:563–591.
30. Travis LB. The epidemiology of second primary cancers. *Cancer Epidemiol Biomarkers Prev*. 2006;15(11):2020–2026.
31. Rothmann KJ, Greenland S, Lash TL. *Modern Epidemiology*. Third ed. Wolters Kluwer: Lippincott Williams & Wilkins; 2008.
32. Hewitt ME, Greenfield S, Stovall E. *From Cancer Patient to Cancer Survivor: Lost in Transition. (Press NA, ed.)*. The National Academies Press; 2006.
33. Ministry of the Interior and Health of Denmark. Law information (In Danish: retsinformation). Available from: https://www.retsinformation.dk/eli/lta/2020/1338. Accessed February 10, 2023.