IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

The Cost of Late Diagnosis: A Synthetic-Data-Based Study of Lung Cancer Survival by Stage

Seo-Yun Jung

Concord College, Acton Burnell, Shrewsbury, Shropshire, SY5 7PF, UK; sharonjung08@gmail.com
Mentor: Grace Jeonghyun Kim

ABSTRACT: The most common cause of cancer mortality in Korea is lung cancer, and patients are usually diagnosed at a late stage. This article examines the association between the diagnostic stage and survival based on a large population representative synthetic electronic health record dataset. A stage-based analysis of survival was performed, followed by an analysis of postoperative outcomes and sex-specific trends of survival. A considerable loss in survival was found between the early (Stages I-II) and locally advanced (Stage III) stages. The differences in survival between Stages III and IV were insignificant, indicating that the inability to resect surgically is a clinically relevant prognostic marker. Sex-stratified tests found the following differences in survival patterns: Women had steeper declines at certain stages, whereas men had a gradual decline. Multivariate analysis suggests that sex is an independent predictor of survival. These results highlight the role of early diagnosis and indicate that sex may play a role in stage-specific survival. The use of synthetic data obviates privacy issues, but it must be validated by real-world clinical data. Finally, this study, in addition to justifying the importance of early diagnosis, indicates areas for future research to improve the countrywide screening and management of lung cancer.

KEYWORDS: Medicine, Epidemiology, Lung Cancer, Stage-Specific Survival, Korean Population.

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, accounting for close to one-fifth of cancer-related deaths across the world.[1] This is a major global public health concern, especially in high-income nations, where its incidence and mortality have been rising in the last few decades. The World Health Organization (WHO, 2025) indicates that the incidence of cancer has increased worldwide, with the number of deaths caused by lung cancer ranking first. Lung cancer is the major cause of cancer-related deaths in South Korea, with an age-standardized rate of incidence of approximately 595.4 in 100,000 men and 543.5 in 100,000 women in 2024.[2] The crude mortality rate of lung cancer is high at 23.0 per 100,000, even with progress in the medical sphere.[3] The main cause of such high mortality is that lung cancer is frequently only diagnosed at its advanced stages, i.e., in Stages III and IV, which significantly lowers the treatment possibilities and survival rates.[4]

Most cases are either locally advanced or distant metastasis.[5] Unlike other types of cancer, which are usually identified early because of particular symptoms or routine screening, lung cancer, in most cases, is not diagnosed until it has spread outside the local region. This leaves limited treatment options and low chances of survival, as the survival rates show a marked decline in survival as cancer advances: The five-year survival rate of non-small cell lung cancer (NSCLC) stands at about 82%, which is Stage I, but is only about 10% in Stage IV.[2] This substantial difference in survival underlines the need to detect lung cancer early enough and hence the need for stronger screening tools to detect the disease, especially among high-risk groups, particularly in South Korea, where the incidence is high and early detection measures are needed to reduce the increasing burden of lung cancer.

Cancer screening programs are important in the detection of cancers at earlier and more treatable stages. The benefits of early detection have already become apparent in the case of other types of cancer, especially cervical, colorectal, and breast cancer. In the United States, cervical cancer rates have declined by about 55 per cent in the last few decades, and a European study found that extensive cervical cancer screening programs reduced mortality rates by 41-9%.[6,7] Likewise, a vast Swedish cohort study showed that the death rate of women with breast cancer has declined among the regularly screened population.[8] These screening programs have already shown the possibility of identifying cancer during its early stages, allowing interventions to be provided earlier, leading to increased survival rates and reduced costs of treatment in the long term.

The same success has not been achieved in screening programs for lung cancer as for cervical, colorectal, and breast cancers; however, despite its high mortality rate, especially at the advanced stages, which, as noted, is the leading cause of cancer-related death among all cancers.[4] Although there is an extensive body of research revealing the benefits of early screening of other types of cancers, the data on survival rates of lung cancer by stage of diagnosis remain limited, especially in Korea. Most research has investigated general cancer occurrence or death without considering the important effect of the diagnostic stage on patient outcomes, leaving us with a lack of information on the stage-specific survival outcomes of lung cancer patients in Korea, particularly in the earlier stages,[9] when the cancer is localized and may be removed via surgical resection or alternative curative interventions.[5]

Staging lung cancer is an important clinical process in which the Tumor Node Metastasis (TNM) classification is employed to determine the progression and prognosis of lung cancer. Stage I and II lung cancer are usually localized and have high survival rates after surgical procedures, but Stage III and IV cancer are more widespread and often require a combination of therapies, including chemotherapy and radiation; the curative treatment options are limited because lung cancer at these stages is typically surgically inoperable, and survival rates are much reduced.[5] More efficient tools for the early detection of lung cancer are thus needed to facilitate diagnosis at the earliest possible stage.

A significant issue in large-population-based research on lung cancer survival has been difficulties in gaining access to extensive patient-level data while protecting patient privacy. Synthetic electronic health record (EHR) data offers one solution to this dilemma. EHR data are generated from de-identified real-world data, which protects patient privacy but still offers a comprehensive picture of patient outcomes. These data can be useful for examining survival trends and other significant health outcomes of large groups of people, particularly in nations like South Korea, where conventional data access may be limited by privacy issues.[10] This study utilized a large synthetic EHR data set to examine the relationship between the diagnostic stage and survival of Korean lung cancer patients. The analysis of this data helps mitigate the limitations of traditional registries, which do not tend to give extensive stage-specific survival data, and allows the results to be applied to the broader population of Korea. Accordingly, this study aimed to address a gap in our knowledge of the stage-specific survival rates of Korean lung cancer patients.

In particular, this study examines the survival times by stage at diagnosis in patients with lung cancer in Korea. We hypothesize that the later the patients are diagnosed, the shorter their survival times, as the stage at diagnosis is anticipated to be a key predictor of survival.[11] In addition, this study examines the possibility of a moderating effect of sex on survival results, as sex-related disparities in lung cancer survival have been reported, such that women generally have a steeper survival curve at certain stages.[4] The results of this study offer insight into the stage-specific survival of lung cancer patients in Korea and their potential implications for health policy to enhance early detection and treatment measures, as well as for the design and evaluation of national screening programs like the Korean Lung Cancer Screening Program (NLCSP), which has demonstrated the effectiveness of early detection, but without full knowledge of the overall survival outcomes.[12]

Methods

This retrospective, observational study used synthetic electronic health record (EHR) data that the South Korean government synthetically created from the country's national cancer registry and tertiary hospital records using artificial intelligence (AI)-driven modeling. In short, the dataset was built from synthesized real-world clinical information that represents the demographic and clinical characteristics of Korean cancer patients but does not contain any identifiable patient information. Using CONNECT, a large healthcare data platform supported by the National Information Society Agency (NIA). The CONNECT platform includes structured clinical data from the colorectal cancer clinical library (converted to lung cancer records) and the national cancer registry (supplemented with tertiary hospital data) that covers most cancer patients in Korea to ensure representativeness at the population level.

The primary outcome was the survival time in months from the date of original diagnosis of lung cancer to the date of death. No censoring was performed because all the subjects in the data were deceased. Stage diagnosis was the primary independent variable that was determined based on Tumor-Node-Metastasis (TNM) guidelines, where T indicates the size and extent of the primary tumor; N indicates the amount of a regional lymph node that is involved; and M indicates the occurrence of distant metastasis. Based on these criteria, a stage (I–IV) was given to every patient. To test the study hypothesis, the four standard stages were categorized as early stage (Stages I–II) or advanced stage (Stages III–IV). There was a list of additional variables that could be used as covariates, not all of which were in the key model: demographic variables, including age, sex, previous treatment history (surgery, chemotherapy, and radiation); comorbid diseases like chronic obstructive pulmonary disease (COPD) and pulmonary fibrosis; and other pertinent clinical outcomes.

R (version 4.5.1) was used for all statistical analyses. Continuous variables were described with means (*M*) and standard deviations (*SD*), whereas categorical variables were described in the form of frequencies (*n*) and percentages (%). Linear regression (OLS) was conducted where the outcome was the survival days (OVRL_SRVL_DTRN_DCNT) and stage at diagnosis (I–IV) was the primary independent variable. OLS was used since all individuals were deceased; hence, the survival period was known for all cases. Although the use of OLS regression in this case might be appropriate, the data are only for deceased subjects, which can limit its representativeness through selection bias by excluding longer-lived subjects, resulting in underestimated survival times and thereby biased stage comparisons. Therefore, the findings reflect only deceased subjects. In this context, OLS regression provides a straightforward approach to estimating mean differences in survival time across groups, as complete survival durations are observed for all individuals. However, this approach differs from conventional survival analysis methods, such as Kaplan–Meier estimation or Cox proportional hazards models, which are typically used when censoring is present. Therefore, the findings should be interpreted with caution, as the analytic framework and sample structure might have influenced the observed associations. Sex (BSPT_SEX_CD) was incorporated as a covariate in the adjusted model, and further models were stratified by sex (male and female). The stage x cancer subtype (BSPT_FRST_DIAG_CD) interaction model was fitted to determine whether the differences in survival based on stage were different across subtypes. A t-test was used to compare the survival between stages (e.g., Stage I vs II/III/IV) and other outcomes (e.g., days from first operation to death, days from first diagnosis to death) across stages, with the results visualized as boxplots, violin plots, and heatmaps. Model diagnostics were conducted through the analysis of the residual distributions in terms of the Q-Q plot of the linear regression models. Two-sided tests were used to determine statistical significance, with $p < 0.05$, and 95% confidence intervals were presented; no multiple testing correction was applied to exploratory *t*-tests.

As this study used entirely synthesized EHR data without recognizable patient data, patient confidentiality was maintained, and institutional review board (IRB) approval and informed consent were not needed. The study complied with the ethics of researching human health records by using synthetic data that is publicly available.

Results and Discussion

We now turn to the findings of our stage-specific survival analyses. The demographic and clinical characteristics of the study population are provided first, followed by the comparison of the survival outcomes by diagnostic stage. A total of 4,307 lung cancer patients were included in the study (Table 1). The mean age at diagnosis was 66.2 years ($SD = 8.8$). Most patients were male (87.9%, $n = 3789$), and only a minority were female (12.0%, $n = 518$), which is in agreement with the fact that Korean men have higher smoking rates and consequently higher lung cancer morbidity rates than Korean women. In terms of stage at diagnosis, most patients presented with advanced-stage disease: Only 19.3% ($n = 832$) of patients presented with Stage I disease, while 41.2% ($n = 1,773$), 22.5% ($n = 969$), and 16.5% ($n = 710$) presented at Stages II, III, and IV, respectively. This indicates that most patients in this cohort were not discovered by early detection and screening, highlighting the need for better lung cancer screening and diagnostic methods.

Treatment type was also influenced by stage. Surgery was most common among early-stage patients, in which curative resection was possible for Stages I–II disease. The small difference in survival between Stages III and IV is notable. Although it may reflect the loss of the possibility for surgical resection at the advanced stage, other factors must be taken into account. They include the heterogeneity of therapeutic strategies, as patients at later stages might receive different forms of treatment such as chemotherapy, radiation therapy, or any other type of treatment again, reflecting the utility and applicability of these treatments in the absence of surgical options. As such, the distribution of stage at diagnosis in this cohort is in line with the natural history of lung cancer and is a useful context for the subsequent survival data.

Table 1: Baseline characteristics of lung cancer patients ($N = 4,307$). The study group consisted predominantly of male patients diagnosed with advanced-stage disease, with more than 80% presenting at Stages II–IV. This pattern indicates a low likelihood of early disease detection in this population.

Variable	Category	Frequency (n)	Proportion (%)	Mean	*SD*
Sex	Male	3789	87.97		
	Female	518	12.03		
Age at diagnosis	-	-		66.22	8.84
Diagnostic stage	I	832	19.32		
	II	1773	41.17		
	III	969	22.5		
	IV	710	16.48		

Survival by Stage:

Survival time decreased as the stage at diagnosis increased (Figure 1). Patients with Stage I disease had the longest survival (mean = 886.1 days; median = 837.5 days). And those with Stage II disease have a slightly shorter survival (mean = 831.8 days; median = 761.0 days). The decline in survival was steeper for patients with Stage III disease (mean = 747.7 days; median = 652.0 days). Unexpectedly, however, the mean survival for Stage IV patients was only marginally longer (mean = 754.0 days; median = 689.5 days) than for Stage III patients, although this difference was not significant. There is a non-linear relationship between stage and survival, and this may be a result of heterogeneity in treatment assignment across patients with advanced disease, differences in patient biology, or palliative effects.

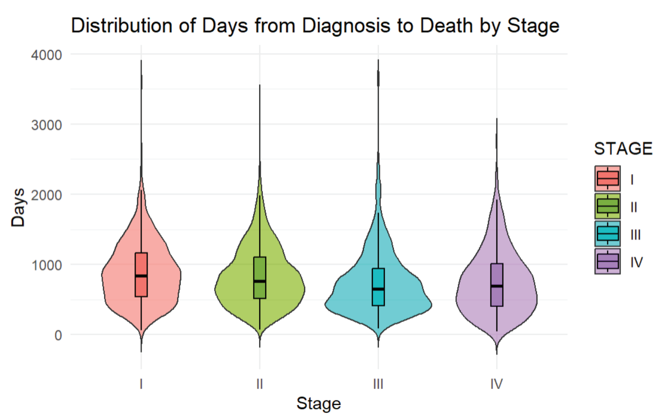

Figure 1: Survival duration (days) from diagnosis to death by diagnostic stage in Korean lung cancer patients. There is a marked difference in survival between Stage I–II disease and Stage III disease, with only a slight difference between Stages III and IV. This trend suggests that the loss of surgical resectability, rather than further disease progression, is a major determinant of survival.

Pairwise comparisons clarified the foregoing stepwise trend (Figure 2). Compared with Stage I disease, median survival decreased significantly at all later stages, by 54.4 days (Stage II, $p < 0.01$), 138.4 days (Stage III, $p < 0.01$), and 132.1 days (Stage IV, $p < 0.01$). Similarly, Stage II patients had longer survival than those with Stage III or IV disease by 84.0 days ($p < 0.01$) and 77.7 days ($p < 0.01$), respectively. By comparison, the difference in median survival between Stages III and IV was minimal (6.3 days, $p = 0.782$), suggesting that once the disease

DOI: 10.36838/IJHSR816.144

is no longer resectable, further progression has little impact. These findings highlight the importance of early detection for survival.

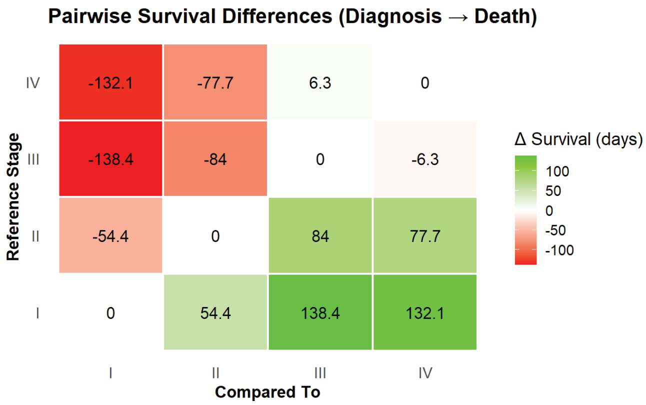

Figure 2: Heat map - diagnosis-death. Survival differences are pronounced when comparing Stage I with Stage IV disease, but are minimal between Stages III and IV. This pattern supports the interpretation that the transition from surgical resectability to non-resectability, rather than further stage progression, is a key determinant of survival.

Post-Operative:

Survival Within the subgroup of patients who were resected, survival progressively decreased at more advanced stages (Figure 3). The average survival days after surgeries were 844.7 (median = 805.5) in Stage I, 795.8 (median = 724.0) in Stage II, 725.3 (median = 627.0) in Stage III, and 711 (median = 645.0) in Stage IV. Regression analysis indicated that surgery performed at earlier stages was significantly associated with greater survival benefits. These results are summarized in the triangular heatmap (Figure 4): Survival declined markedly from Stages I–II to Stages III–IV. Collectively, these data highlight the need for curative resection before local or systemic metastasis and indicate that the surgical prognosis is associated with poorer outcomes once the disease is no longer resectable.

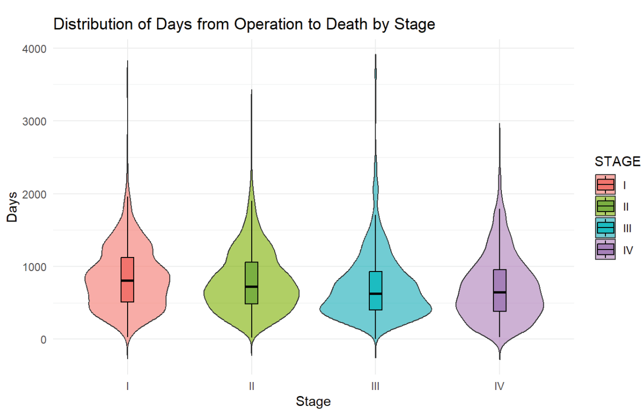

Figure 3: Post-operative survival duration (days) from operation to death by diagnostic stage. Post-operative survival decreases progressively with advancing stage at diagnosis, with the greatest survival observed in patients who undergo surgery at early stages. These findings suggest that the survival benefit of surgery diminishes once tumor burden exceeds resectable limits.

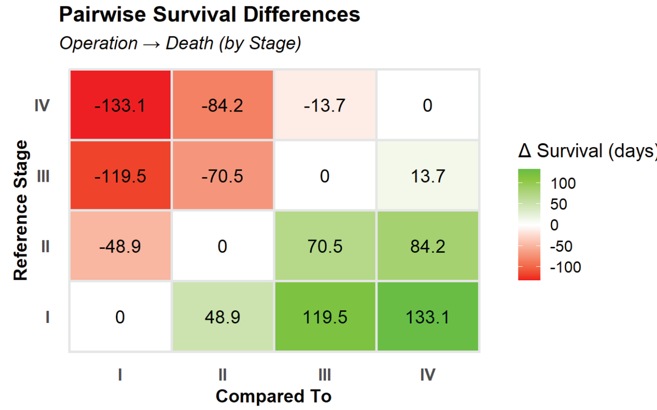

Figure 4: Heat map - operation-death. Large survival advantages are observed for patients undergoing surgery at earlier stages, whereas post-operative survival differences between Stages III and IV are minimal. This pattern indicates that the survival benefit of surgery declines once the disease exceeds resectable limits.

Sex Effects on Survival:

Sex-stratified analysis showed differences in stage-specific survival (Tables 2 and 3), such that survival decreased sharply in women at stages II (–179.8 days, $p < 0.001$) and again at IV (–327.0 days, $p < 0.001$) compared to stage I, with a non-significant decline at III (–181.8 days; $p = 0.165$). These findings suggest that female patients may lose their survival benefit more rapidly during the first phase, possibly at least in part due to underlying biological differences, tumor behavior, or response to therapy.

Male patients demonstrated a more stepwise and regular decrease: Survival was reduced by –26.9 days at Stage II ($p = 0.178$), –84.4 days at Stage III ($p < 0.001$), and –102.6 days at Stage IV ($p < 0.001$). This more stable trend is thought to be due to the greater underlying disease burden because of longer periods of smoking-induced pathology among Korean men.

The model was further adjusted for sex and cancer subtype to examine the interactions between these patient characteristics. Although there was a trend of decreasing survival with increasing stage, the effect of cancer stage did not reach statistical significance when adjusted for other prognostic factors for the disease. Compared to Stage I, Stage II survival was lower by –222.9 days (p= 0 .343), by –263.8 days for Stage III ($p = 0.168$), and by –112.0 days for Stage IV ($p = 0.402$). Unlike stage, sex was also an independent predictor for survival in the multivariate analysis: Men had a predicted survival 116 days shorter than women did on average ($p < 0.001$). Between cancer subtypes, minimal nonsignificant differences were found ($p > 0.1$ for all).

These results suggest that sex is a significant effect modifier for stage-specific survival in Korean lung cancer patients. For women at specific stages of disease, survival declines more steeply; for men, it is a more gradual decline. The sex difference in survival did not vary by cancer subtype and may need to be considered in lung cancer screening, staging, and treatment.

Table 2: Sex-stratified regression of survival duration by stage at diagnosis in Korean lung cancer patients. Stage-specific survival patterns differ by sex, with women exhibiting sharper declines in survival at selected stages, while men show a more gradual reduction with advancing stage. These findings indicate that sex modifies the relationship between diagnostic stage and survival.

	Female		Male	
Stage	Survival Deviation from Stage 1	p-value	Survival Deviation from Stage 1	*p*-value
II	-179.8	<0.001	-26.9	0.178
III	-181.8	0.165	-84.4	<0.001
IV	-327	<0.001	-102.6	<0.001

Table 3: Sex-adjusted regression of survival duration controlling for cancer subtype. After adjustment for diagnostic stage and cancer subtype, sex remains an independent predictor of survival, whereas stage effects are reduced. This suggests that sex-related factors contribute to survival differences beyond disease stage alone.
Diagnosis codes correspond to ICD-10 classifications for lung cancer (C34.0–C34.9).

Variable	Estimate	P-value	97.5 % Confidence Interval (Lower–Upper)
Stage II	-222.86	0.343	-461.06 to +15.34
Stage III	-263.79	0.168	−374.88 to −152.70
Stage IV	-111.97	0.402	−261.79 to +37.85
Sex (female: control)	-116.04	<0.001	−191.44 to −40.64
ICD-10 C34.1	7.21	0.94	−189.38 to +203.80
ICD-10 C34.2	-158.46	0.122	−200.59 to −116.33
ICD-10 C34.3	-76.34	0.43	−189.78 to +37.10
ICD-10 C34.8	52.37	0.821	−452.53 to +557.27
ICD-10 C34.9	-21.53	0.825	−191.07 to +148.01

* ICD-10 C34.1: Malignant neoplasm of the upper lobe, bronchus, or lung
* ICD-10 C34.2: Malignant neoplasm of the middle lobe, bronchus, or lung
* ICD-10 C34.3: Malignant neoplasm of the lower lobe, bronchus, or lung
* ICD-10 C34.8: Overlapping lesion of the bronchus and lung
* ICD-10 C34.9: Malignant neoplasm of the unspecified part of the bronchus or lung

In this nationwide, large Korean cohort of 4,307 lung cancer patients, the majority of whom were male, most patients were diagnosed at an advanced stage, consistent with limited early detection. Survival was reduced from stage to stage with a marked decline between early-stage (I–II) and locally advanced-stage (III) and a nearly flat slope between stages III and IV. There was a significant survival advantage in patients who underwent surgery at an early stage, highlighting the importance of early diagnosis and management. There were differences by sex in stage-specific lung cancer survival, however. Women had a steeper decline in survival at some stages, while a less pronounced decrease was found among men, and sex was an independent predictor of survival adjusted for stage and subtype, with males living an average of 116 days shorter than females ($p < 0.001$). These findings suggest that stage at diagnosis and sex were key factors associated with overall survival among Korean patients with lung cancer.

This study sets out to test the association between stage at diagnosis and survival time among Korean lung cancer patients based on a large-scale synthetic electronic health record (EHR) dataset. The results provide insights into the stage-specific survival outcome of lung cancer in Korea, the significance of early diagnosis, and the critical role of the stage at diagnosis in terms of long-term survival. Also, the research highlights the impact of sex on survival rates, suggesting that sex-specific treatment could help improve lung cancer survival.

Effect of Stage of Diagnosis on Survival:

The findings of this study are in line with those of previous studies, which indicated that the duration of survival declines substantially as the diagnosis stage progresses.[9] Patients in Stage I had the highest survival times, with a mean survival time of 886.1 days. This result highlights the prognostic benefit of early-stage diagnosis and is consistent with the worldwide knowledge of the prognostic value of cancer at an early stage that is localized and can be treated with surgery, giving the highest possibility of survival.[11] Survival of Stage II patients was relatively good; however, there was a marked decrease from that of Stage I, and a mean survival of 831.8 days was observed. Nevertheless, the strongest outcome was observed when the comparison was made between Stages III and IV. Although it may be reasonable to assume that there would be a greater difference in outcomes between these stages, the results showed that there was minimal difference in survival of patients in these two stages, with a mean survival time of about 754.0 days in Stage IV and 747.7 days in Stage III. This implies that when the disease is at Stage III, additional development to Stage IV was not associated with a substantial difference in the chances of survival. This aligns with other research results that show that when the disease has reached local progression and is no longer controllable, the prognoses do not improve with further progression of the disease.[4]

Sex Effects on Stage-Specific Survival:

One of the notable findings of this study was the sex-specific differences in the survival trends. The survival of women declined markedly in Stage II, and again in Stage IV, but there was a slower decline in men in the stages. The findings are consistent with those of the earlier findings of better survival outcomes in women at various stages of lung cancer, potentially because of biological differences, including hormonal or tumor biological variations.[13] Conversely, the decline in survival was more linear among men, who smoke more and, therefore, have a greater chance of encountering lung cancer.[4] The identified sex differences in survival results indicate that sex-specific approaches to lung cancer diagnosis, therapy, and screening are necessary. Multivariate analysis showed that sex remained an independent predictor of survival, where men lived an average

DOI: 10.36838/IJHSR816.144

of 116 days shorter than women, despite the cancer stage and subtype. The results can be used to support the hypothesis that biological and behavioral disparities between men and women influence survival outcomes and that sex-specific interventions can help to achieve better outcomes in terms of survival.[13] These findings are important in informing the optimal public health policies, which might help in establishing more individualized policies on lung cancer detection and treatment.

The Rationale of Early Detection and National Screening Programs:

The results in this research indicate that earlier diagnosis is important in helping lung cancer patients have a better chance of survival. As Stage I patients were found to be associated with substantially higher survival rates, there is an indication that early diagnosis is an important factor in prolonging the life span. The information also indicates that the difference between Stage I and Stage III is large, whereas the difference between Stage III and Stage IV is comparatively minor. This shows that the greatest survival advantages are achieved when the disease is detected before reaching a stage where curative treatment is no longer feasible. This finding offers support in favor of advancing the lung cancer screening programs, including the Korean Lung Cancer Screening Program (NLCSP), which has suggested the potential to identify early-stage cancers. The effectiveness of screening programs for other types of cancer, like cervical and colorectal cancer, further supports the need to implement cancer screening programs on a population level.[7] There are examples of programs that focus on early detection of cancer that have been credited with greatly lowering mortality of people in several countries, and a similar program targeting lung cancer may contribute to reductions in mortality.[6] Although the initial outcomes of the Korean Lung Cancer Screening Project (K-LUCAS) are promising, the results require confirmation using the survival data by stage to optimize these programs for national implementation.[14]

Limitations of Synthetic EHR Data:

Although there are several benefits to the synthetic EHR data set utilized in this research, especially regarding privacy issues and the delivery of population-wide information, it has several limitations that should be borne in mind. Although synthetic data are helpful in a large-scale epidemiological study, they might be inadequate in reflecting the heterogeneity of clinical data in the real world, especially in terms of treatment procedures and patient behavior. As an example, the data employed in the present study did not include such variables as smoking history, socioeconomic status, and comorbidities, which have also been reported to influence lung cancer survival.[15] In addition, the synthetic dataset does not have the same clinical detail as real-world data, so it is not generalizable to actual populations. However, synthetic data can provide information about stage-specialized survival with greater accessibility and replicability. Another limitation is that the synthetic EHR data do not provide detailed treatment data. The choice of treatment modalities, i.e., chemotherapy, surgery, and radiation, is also a key determinant of survival outcome among patients with lung cancer.[16] Further research based on actual clinical information must incorporate treatment-related variables for a better picture of the factors underlying the difference in survival rates between stages. Also, despite the differences in survival rate between men and women, there remains an obvious bias in the population, with more men (88%) than women (12%). As a result, the validity of sex-based survival difference comparisons might be questionable because of the large difference in the number of patients by sex, which can pose problems for statistical analysis. The specific mechanisms involved were also not explored in this study. Thus, any explanations that attribute these effects to hormone levels or biological processes can be considered hypothetical. A final limitation of the methodology was the use of synthetic data derived by converting records of colorectal cancer to records of lung cancer. Although this approach facilitates research in the absence of actual data on lung cancer, the resulting data may not represent the specific biological and clinical features of patients with lung cancer. This implies that some of the findings of the current method may not be entirely reflective of actual clinical settings.

Future Research:

Directions The present research contributes to research on lung cancer survival, yet substantial gaps remain. Future research should aim to confirm these results with real-world clinical data, which would overcome the limitations of synthetic data and create a more accurate image of the outcomes of lung cancer survival. Specifically, this will be necessary to determine how other factors, including treatment modalities and comorbidities, can affect survival, as such factors are likely to interact with stage at diagnosis and sex to influence results. Also, further research is needed to clarify the mechanisms underlying the sex-specific difference in survival found in this study. Knowledge of the biological, hormonal, and behavioral mechanisms that underlie these differences may inform the development of more targeted treatment strategies and enhance survival outcomes.

■ Conclusion

A population-representative synthetic electronic health record dataset was used in this study to explore the impact of stage diagnosis on the prognosis of Korean lung cancer patients. The survival in the cohort declined between early (Stages I, II) and locally advanced disease (Stage III), with little difference observed between Stages III and IV. The minimal differences in the results after the disease passes the threshold of surgical resectability are an indication that this pattern may reflect the impact of delayed diagnosis, once curative surgical options may no longer be feasible. In practice, the benefits of survival may be expected not in cases of aggressive treatment of the disease at a very advanced stage, but in the diagnosis of the disease at an earlier stage, so that more patients may receive treatment while surgical resection remains feasible.

Post-operative analyses supported these findings. In patients with surgery, survival was decreased with advancing stages, and in those who had surgery in Stages I–II, clear differences in survival were observed from those who received surgery after a delay in Stages III or IV. These findings highlight the importance of diagnosis at an early stage and timely curative resection and lend credence to the argument for enhancing national responses that shift patients left on the diagnostic timeline through improved assessment of risk, earlier referrals, and organized screening of high-risk populations.

Sex-stratified models indicate that stage does not work in isolation. Thus, sex is associated with differences in outcomes, as the abrupt loss of survival in women occurred at certain stages, whereas men experienced a more gradual loss. After adjusting for stage and cancer subtype, sex was an independent predictor of survival, such that men had shorter survival times on average than women. This trend suggests biological, hormonal, behavioral, or treatment-based differences between men and women that can alter how the stage is translated into survival time. This implies that generalized screening and management strategies might overlook the distinct sex-specific risks and intervention opportunities, which can affect the general health of the population and the health care delivered to its members.

The availability of computer-generated EHR data enabled this study to provide detailed, stage-specific survival estimates on a national scale that do not place patient privacy at risk. At the same time, however, synthetic data cannot be entirely substituted for real clinical data. Smoking history, socioeconomic status, comorbidity burden, and detailed treatment pathways are important elements that are not fully represented in the dataset, and the absolute figures generated by the present study must be viewed with some reservations. Future research is needed to validate these findings with data from real clinical registries and incorporate more variables in the analysis to enhance our understanding of the interactions of stage, sex, treatment, and background risk in practice.

Overall, this study suggests that stage at diagnosis and sex are both key factors of survival in Korean lung cancer patients and that the survival declines most markedly after the disease has reached the later stages, when it cannot be surgically removed. It gives quantitative support to the case for earlier detection, more decisive application of curative surgery while the disease remains resectable, and more subtle sex-specific screening and management approaches. Simultaneously, it identifies unresolved issues: the necessity to test synthetic-data results, the necessity to incorporate behavioral and social risk factors into the survival models more thoroughly, and the necessity to refine how national lung cancer screening programs can be fine-tuned to detect more cancers before the disease progresses beyond treatable stages.

Acknowledgments

I would like to sincerely thank Grace Jeonghyun Kim for her mentorship, guidance, and valuable feedback throughout the development of this research.

References

1. Bray F, Laversanne M, Sung H, Ferlay J., *et al.*, *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA: a cancer journal for clinicians, **2024**. 74(3): p. 229-263.
2. Jeong SH, Chun SM, Lee H, Kim M., *et al.*, *Impact of diagnosis-to-treatment interval on mortality in patients with early-stage breast cancer: a retrospective nationwide Korean cohort*. BMC Women's Health, **2025**. 25(1): p. 247.
3. Jani CT, Kareff SA, Morgenstern-Kaplan D, Salazar AS., *et al.*, *Evolving trends in lung cancer risk factors in the ten most populous countries: an analysis of data from the 2019 Global Burden of Disease Study*. EClinicalMedicine, **2025**. 79.
4. Lee JG, Kim HC, Choi CM., *Recent trends of lung cancer in Korea*. Tuberculosis and Respiratory Diseases, **2021**. 84(2): p. 89.
5. Jeon DS, Kim HC, Kim SH, Kim TJ., *et al.*, *Five-year overall survival and prognostic factors in patients with lung cancer: results from the Korean Association of Lung Cancer Registry (KALC-R) 2015*. Cancer Research and Treatment: Official Journal of Korean Cancer Association, **2022**. 55(1): p. 103-111.
6. Adegoke O, Kulasingam S, Virnig B., *Cervical cancer trends in the United States: a 35-year population-based analysis*. Journal of women's health, **2012**. 21(10): p. 1031-1037.
7. Cardoso R, Guo F, Heisser T, Hackl M., *et al.*, *Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study.* The Lancet Oncology, **2021**. 22: p. 1002-1013
8. Mao, Z, Nyström, L, Jonsson, H., *Breast cancer screening with mammography of women 40-49 years in Sweden*. **2014**, Umeå universitet.
9. Lee, J. G., Kim, H. C., Choi, C. M. *Recent Trends of Lung Cancer in Korea*. Tuberculosis and respiratory diseases, **2021**. 84(2), p. 89–95.
10. Yan C, Zhang Z, Nyemba S, Li Z., *Generating synthetic electronic health record data using generative adversarial networks: Tutorial*. Jmir Ai, **2024**. 3: p. e52615.
11. Ziora P, Skiba H, Kiczmer P, Zaboklicka N., *et al.*, *Ten-Year Observational Study of Patients with Lung Adenocarcinoma: Clinical Outcomes, Prognostic Factors, and Five-Year Survival Rates*. Journal of Clinical Medicine, **2025**. 14(8): p. 2552.
12. Kim H, Kim HY, Goo JM, Kim Y., *Lung cancer CT screening and Lung-RADS in a tuberculosis-endemic country: the Korean Lung Cancer Screening Project (K-LUCAS)*. Radiology, **2020**. 296(1): p. 181-188.
13. Radkiewicz C, Dickman PW, Johansson ALV, Wagenius G., *et al.*, *Sex and survival in non-small cell lung cancer: A nationwide cohort study.* PloS one, **2019**. 14(6): p. e0219206.
14. Kim Y, Lee J, Lee E, Lim J., *et al.*, *Strategies to improve smoking cessation for participants in lung cancer screening program: analysis of factors associated with smoking cessation in Korean Lung Cancer Screening Project (K-LUCAS)*. Cancer Research and Treatment: Official Journal of Korean Cancer Association, **2024**. 56(1): p. 92-103.
15. Avci T, Sahbazlar M, Ekinci F, Erdogan AP., *The Advanced Lung Cancer Inflammation Index: A Powerful and Independent Prognostic Factor in Metastatic Gastric Cancer*. Medicina, **2025**. 61(11): p. 1949.
16. Alexander M, Wolfe R, Ball D, Conron M., *et al.*, *Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer*. British journal of cancer, **2017**. 117(5): p. 744-751.
17. World Health Organization (2019). ICD-10: International Statistical Classification of Diseases and Related Health Problems. Available at: https://icd.who.int/browse10/2019/en#/C34

 DOI: 10.36838/IJHSR816.144

Author

I am a Year 12 student studying mathematics, biology, and chemistry, and I aspire to study medicine at university. I have strong motivation for this field due to my ambition to combine clinical practice with research to improve diagnosis, treatment, and health policy.

DOI: 10.36838/IJHSR816.144