The Clinicopathological Features and Survival Outcomes of Patients with Different Metastatic Sites in Stage IV Endometrial Cancer:A SEER-Based Cohort Study

Mohammed Zuber

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). Adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

**Authors**

* Mohammed Zuber (ORCID: 0000-0000-1234-5678)

**Author affiliations**

1. College of Pharmacy, University of Georgia, Athens, GA, USA.

Corresponding author: mohammedzuber099@gmail.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis* Endometrial cancer, a neoplasm originating from the endometrial lining, is the most common gynecological cancer in the United States. According to American Cancer Society 2024 statistics the incidence of uterine cancer is 67, 880 with 13,250 deaths. Due to its early symptomatic presentation, 70% of cases are diagnosed at stage I, where the five-year survival rate exceeds 90%. However, around 13% of patients are diagnosed at advanced stages (III/IV).

A retrospective study found that the prevalence of stage IVB endometrial cancer increased to 5.67% between 2010 and 2015. Unlike stage I, which has a favorable prognosis, stage IV is associated with poor outcomes, with a five-year survival rate ranging from 17% to 30%. The most frequent metastatic sites include the peritoneum, lungs, pelvic and para-aortic lymph nodes, and local pelvic recurrences. Among these, lung metastasis is one of the most common distant metastases, occurring in 1.5% of patients. The hematogenous spread is the primary mechanism of lung metastasis. Depending on clinical features, disease stage, prior surgical history, and pathological factors, 11-13% of endometrial cancer patients experience recurrence within the first two years after initial treatment. Symptoms of lung metastasis typically include coughing up blood, shortness of breath, chest pain, persistent cough, pleural effusion, reduced appetite, and weight loss.

Compared to other gynecological cancers such as cervical and ovarian cancer, endometrial carcinoma exhibits the highest rate of pulmonary metastasis, affecting up to 20-25% of patients who experience recurrence. According to Mao et al., 2020, the five-year cancer-specific survival (CSS) rate for liver and brain metastases was 21%, while bone metastases had a survival rate of 20%. In patients with lung metastases, the median overall survival was 11 months, with a five-year CSS rate of 19%. However, this data specifically pertains to stage IV lung cancer patients.

Patients with advanced endometrial cancer face poor survival outcomes, with metastatic disease being the leading cause of mortality. Therefore, this study aimed to utilize data from the Surveillance, Epidemiology, and End Results (SEER) database to analyze survival outcomes in patients diagnosed with stage IV endometrial cancer. The primary objective was to determine whether the presence of lung/brain/bone/liver metastases in adults with stage IV endometrial cancer is associated with worse survival outcomes compared to patients without any metastasis.

Please note the references are not yet added and will be added soon.

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.* This study will utilize the Surveillance, Epidemiology, and End Results (SEER) database, one of the most comprehensive cancer incidence and survival registries, covering over 48% of the United States. SEER collects key demographic data, cancer characteristics, treatment details, and survival outcomes. Managed by the U.S. National Cancer Institute, the database began recording information on lung, bone, liver, and brain metastases in 2010. Data for this study will be accessed through SEER\*Stat 8.4.4 software, using the SEER Research Data (2000–2019) from 17 registries.

The extracted variables include age, race, income, American Joint Committee on Cancer (AJCC) stage, year of cancer diagnosis, lung metastasis (yes/no), brain metastasis (yes/no), liver metastasis(yes/no), brain metastasis (yes/no), cause of death, survival months, vital status (dead/alive), rural/urban classification.

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

*Question* What are the clinicopathological features and survival outcomes of patients with different metastatic sites in stage IV endometrial cancer?

*Hypothesis* Patients with stage IV endometrial cancer exhibit distinct clinicopathological characteristics and survival outcomes based on the site of metastasis. Specifically, metastases to different organ sites (e.g., lung, liver, bone, and brain) are associated with varying prognostic implications, with some metastatic patterns demonstrating significantly poorer survival compared to others. By analyzing data from the SEER database, we hypothesize that certain metastatic sites will be linked to more aggressive tumor behavior and worse overall survival, highlighting potential differences in disease progression and treatment response.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (1,2).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.* This study examines survival outcomes in stage IV endometrial cancer using SEER data. The primary outcomes include survival months, cause-specific death classification, and vital status at study cutoff. Key predictors include age, race, income, year of diagnosis, and presence of differnt metastasis status.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.* The data was extracted from SEER\*Stat 8.4.4 software, using the SEER Research Data (2000–2019) from 17 registries. The following inclusion criteria was used to extract the data (1) primary site, C54.1–9 and C55.9 based on International Classification of Tumor Diseases for Oncology (ICD-O) codes; (2) year of diagnosis 2010-2015; (3) AJCC stage IV; (4) age more than 18 years.

The SEER database provided data on both demographic and clinical characteristics of the patients. The demographic variables included age at diagnosis, race, median household income (inflation-adjusted to 2019), year of diagnosis, and Derived AJCC Stage Group (7th edition, 2010-2015). The clinical variables comprised lung, brain, bone, and liver metastasis status, survival duration (in months), SEER cause-specific death classification, and vital status recode (used for study cutoff determination).

Age will be categorized into four groups: 18-39, 40-59, 60-79, and ≥80 years. Race is classified into Asian or Pacific Islander, American Indian or Alaska Native, Black, and White. The income variable will be divided into six brackets: < $35,000, $35,000 - $44,999, $45,000 - $54,999, $55,000 - $64,999, $65,000 - $74,999, and $75,000+. The SEER cause-specific death classification was preferred over all-cause mortality to ensure that only deaths specifically attributed to endometrial cancer will be included in the analysis.

## 3.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

Endometrial cancer cases will be extracted from the SEER database. Patients with an unknown lung/brain/bone/liver metastasis status will be excluded. Individuals under 18 years of age will be removed from the dataset. Additionally, patients whose cause of death is unrelated to endometrial cancer or whose cause of death is missing or unknown will not be included in the analysis. Cases diagnosed at AJCC stages I to III will be excluded from the study. The final cohort will be categorized into 5 groups: lung metastasis, brain metastasis, bone metastasis, liver metastasis and no metastasis.

The data cleaning process has not yet started. The R code will be provided as the analysis completed.

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.* The analysis aims to identify survival differences based on metastatic patterns, demographic disparities, socioeconomic influences, and temporal trends in prognosis. Kaplan-Meier survival analysis will used to compare the survival among the study cohorts. The incidence of metastasis will be calculated in stage IV endometrial cancer. A Cox regression model will be performed to compute the adjusted Hazard Ratios (aHR) and 95% confidence intervals (CI).

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (3) discusses types of analyses.

These papers (1,2) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

1. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

2. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.

3. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.