MSDS 570 Final Project

Andrea Hannah

MMC

School of Applied Computational Sciences

[handrea22@email.mmc.edu](mailto:handrea22@email.mmc.edu)

***Abstract***— Identifying biomarkers that predict when patients are at risk for ovarian cancer, the fifth deadliest cancer for women, is the key to improved survival rates. Ovarian Cancer is a group of diseases that originate in the ovaries or in the fallopian tubes or peritoneum. Ovaries produce the female hormone and reproductive eggs. Ovarian cancer is best treated at its earliest stages before becoming aggressive. Therefore, early screening and diagnosis is key to successfully treating the disease, curing, or entering remission. Using the heatmap method of exploratory data analysis to assess correlations and using existing literature, the biomarkers of importance identified include Age, Menopause, Human Epididymis Protein 4 (HE4), Alkaline Phosphatase (ALP), and Calcium. Using Scatterplot visualizations, each contributing variable is assessed for comparison with elevated CA125 levels. All variables of interest, except HE4, correspond with elevated CA125 levels and would be biomarkers to play closer attention to in ovarian cancer screening.

***Keywords—Ovarian Cancer; Gynecological Health; Cancer Antigen; Big Data Visualizations; Gynecology; HE4; ALP; Ca; CA125***

1. INTRODUCTION

Ovarian Cancer is a group of diseases that originate in the ovaries or in the fallopian tubes or peritoneum. Ovaries produce the female hormone and reproductive eggs. Ovarian cancer is best treated at its earliest stages before becoming aggressive, as it is the fifth deadliest cancer to women. Therefore, early screening and diagnosis is key to successfully treating the disease, curing, or entering remission. Ovarian Cancer is screened through two methods:

1. TVUS (transvaginal ultrasound, a test using sound waves to look at the uterus, fallopian tubes, and ovaries. It can find a mass, but it can't tell if a mass is cancer or benign. When it is used for screening, most of the masses found are not cancer.
2. The CA-125 blood test measures the amount of a protein called CA-125 in the blood. Many women with ovarian cancer have high levels of CA-125. This test can be useful as a tumor marker to help guide treatment in women known to have ovarian cancer, because a high level often goes down if treatment is working.

Checking CA-125 levels has not been found to be as useful as a screening test for ovarian cancer because high levels of CA-125 is more often caused by common conditions such as endometriosis and pelvic inflammatory disease, and, not everyone who has ovarian cancer has a high CA-125 level.

Therefore, other biomarkers should be used in conjunction with CA-125 to screen for ovarian cancer.

# II. DATASET AND FEATURES

The original Ovarian Cancer Dataset has 349 observations. There are 51 variables including Patient ID, Age, Menopause (Y/N), and Type, as well as 47 biomarkers found in blood or serum samples. This dataset includes the following biomarker features:

|  |  |  |
| --- | --- | --- |
| Abbreviation | Biomarker Name | Sample type |
| AFP | alpha-fetoprotein | Serum |
| AG | Anion gap | Serum |
| ALB | Albumin | Serum |
| ALP | Alkaline phosphatase | Serum |
| ALT | Alanine aminotransferase | Serum |
| AST | Aspartate aminotransferase | Serum |
| BASO# | Basophil Cell Count | full blood |
| BASO% | Basophil Cell ratio | full blood |
| BUN | blood urea nitrogen | Serum |
| Ca | Calcium | Serum |
| CA125 | Carbohydrate antigen 125 | Serum |
| CA19-9 | Carbohydrate antigen 19-9 | Serum |
| CA72-4 | Carbohydrate antigen 72-4 | Serum |
| CEA | Carcinoembryonic antigen | Serum |
| CL | Chlorine | Serum |
| CO2CP | carbon dioxide-combining Power | Serum |
| CREA | Creatinine | Serum |
| DBIL | direct bilirubin | Serum |
| EO# | eosinophil count | full blood |
| EO% | eosinophil ratio | full blood |
| GGT | Gama glutamyl transferase | Serum |
| GLO | Globulin | Serum |
| GLU. | Glucose | Serum |
| HCT | Hematocrit | full blood |
| HE4 | human epididymis protein 4 | Serum |
| HGB | Hemoglobin | full blood |
| IBIL | Indirect bilirubin | Serum |
| K | Kalium | Serum |
| LYM# | lymphocyte count | full blood |
| LYM% | lymphocyte ratio | full blood |
| MCH | Mean corpuscular hemoglobin | full blood |
| MCV | mean corpuscular volume | full blood |
| Mg | magnesium | Serum |
| MONO# | mononuclear cell count | full blood |
| MONO% | monocyte ratio | full blood |
| MPV | Mean platelet volume | full blood |
| Na | Natrium | Serum |
| NEU | neutrophil ratio | full blood |
| PCT | thrombocytocrit | full blood |
| PDW | Platelet distribution width | full blood |
| PHOS | phosphorus | Serum |
| PLT | platelet count | full blood |
| RBC | Red blood cell count | full blood |
| RDW | red blood cell distribution width | full blood |
| TBIL | total bilirubin | Serum |
| TP | Total protein | Serum |
| UA | uric acid | Serum |

The cleaned dataset will leave us with 235 patient observations and 49 features. To begin to understand the cleaned dataset, we created a Heatmap to explore the correlations of the aforementioned variables. The heatmap shows which variables have correlations.

A picture containing table

Description automatically generatedFigure 1

With any research using this dataset, we would drop the highly correlated features shown in this heatmap from the model. PCT and PLT were dropped.

Through exploratory data analysis, a relationship between PCT (thrombocytocrit) and PLT (platelet count) is visible. Age and Menopause share an expected correlation. PCT and PLT are measures of essentially the same thing through different methods, so these variables will be dropped.

Chart, scatter chart

Description automatically generated

Figure 2

It appears that elevated CA125 levels show at age 59 for this dataset. The prevailing notion that ovarian cancer shows up for women 63 years or older is dangerous and it is potentially a factor in late screening and diagnosis.

Graphical user interface

Description automatically generated with medium confidence Figure 4

When looking at the relationship between Menopause and CA125, there is a relationship of higher CA125 levels with the presentation of menopause. However, there may be some outliers, as low CA125 values are also represented with the presence of menopause and vice versa when the onset of menopause has not begun.

Chart, scatter chart

Description automatically generated

Figure 5

Finally, human epididymis protein 4 (HE4) has an inverse relationship than expected. High levels of HE4 are not associated with higher levels of CA125 as expected. Rather, lower levels would have a stronger connection if not for the fact that there are so few scatter plot points as HE4 increases, so that would be disputable and hopefully, proved otherwise with a larger dataset. Scientists have stated initial promise for HE4 in early ovarian detection, but this dataset cannot support it at its present size.

Chart, scatter chart

Description automatically generated Figure 6

In examining the relationship between Calcium with CA125, notice the elevation in CA125 corresponds with an increase in Ca.

Chart, scatter chart

Description automatically generated Figure 3

In examining the correlation between Alkaline Phosphatase with CA125, notice the elevation in CA125 corresponds with an increase in ALP.

III. BRIEF VISUALIZATION METHODS RESULTS

The Heatmap visualization shows which variables are correlated with one another. As a result, I removed PCT and PLT from my variables of interest.

The Scatterplot visualization is used to show the relationship between variables. Using this method, the visualization results yielded:

1. Higher Calcium levels are associated with higher CA125 levels of risk.
2. Higher ALP levels are associated with higher CA125 levels of risk.
3. Higher HE4 levels are not supported as related with higher CA125 levels of risk with this dataset.
4. As expected, menopause and age correlate with higher CA125 levels.

IV. CONCLUSION AND FUTURE WORK

The dataset is smaller, but the potential is expansive and would best be established as a study at Meharry. A future study would be to analyze and predict comorbidities in order to predict the composition of the wholistic care team of specialists (heart-renal-metabolic, etc.) needed that may be required for particular cancer patients to improve survival rates by starting with a larger dataset requested from the National Cancer Institute.

The below missing features could improve research and help with eliminating disparities in gynecological cancers, such as ovarian.

* Demographics
* Lifestyle of mother
* Results of non-stress test
* Previous pregnancies or stillbirths of mother

In addition to the above, increasing the number of observances, i.e. patients, in our dataset will help improve model performance in identifying the most important biomarkers. A primary motivation is to eliminate disparities in screening/diagnosis, treatment, and survival rates for ovarian cancer patients. Unfortunately, the dataset is not as large, however, another motivation is to collect more data by establishing as an official study at Meharry, given that 5-year relative survival rate for ovarian cancer increased from 33% to 48% among non-Hispanic White women but decreased from 44% to 41% in African American women. The study would be a joint collaboration between Meharry institutes: the Center for Advanced Scientific Computing, Innovation and Center of Women’s Health, and eventually the Center of Health Policy since there are disparities in ovarian screening and treatment costs under Medicaid.

REFERENCES

1. February 23, 2023, January 20, 2023, & March 9, 2023. (n.d.). Ovarian cancer studies aim to reduce racial disparities. National Cancer Institute. Retrieved March 31, 2023, from https://www.cancer.gov/news-events/cancer-currents-blog/2020/ovarian-cancer-racial-disparities-studies
2. Ovarian - datasets - PLCO - the cancer data access system. National Cancer Institute - Cancer Data Access System. (n.d.). Retrieved March 31, 2023, from https://cdas.cancer.gov/datasets/plco/23/
3. Shahane, S. (2021, February 6). Predict ovarian cancer. Kaggle. Retrieved March 31, 2023, from <https://www.kaggle.com/datasets/saurabhshahane/predict-ovarian-cancer>
4. American Cancer Society. (n.d.). Can Ovarian Cancer Be Found Early? Early, Detection, Diagnosis, and Staging. Retrieved from https://www.cancer.org/cancer/ovarian-cancer/detection-diagnosis-staging/detection.html