The patient, diagnostic, and treatment intervals in adult patients with breast cancer from high- and lower-income countries

The data source used in this project is from a systematic review and meta-analysis study named "The patient, diagnostic, and treatment intervals in adult patients with cancer from high and lower-income countries: a systematic review and meta-analysis" which can be downloaded from the Open Science Framework: DOI 10.17605/OSF.IO/REY9C (<https://osf.io/rey9c/>).

The research question aims to understand the health disparities among adult patients with breast cancer in high-income and lower-income countries, with a focus on comparing the duration of the patient, diagnostic, and treatment intervals. The dataset used for this analysis is a filtered DataFrame breast\_df obtained from a larger original DataFrame df. The breast\_df contains rows where the 'site\_specific' column is either 'PA breast' or 'breast'. The dimensions of breast\_df are (184, 60), indicating 184 rows and 60 columns. Key input features include demographic details such as age and gender, alongside diagnostic and treatment-related variables like the duration of intervals between diagnosis and initial treatment initiation. Additionally, socio-economic indicators such as Gross National Income (GNI), Human Development Index (HDI), and various aspects of healthcare system performance are incorporated. The outcome of interest primarily revolves around understanding the disparities in the duration of patient, diagnostic, and treatment intervals across different income brackets and countries.

Major findings:

1. Relative contribution of intervals

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Median Ratio | Ratio Median 1 | Ratio Median 2 | Lower Bound | Upper Bound |
| Diagnostic/Patient | 0.932075 | 24.7 | 26.5 | 7.870688e-15 | 1.103798e+14 |
| Diagnostic/Treatment | 0.968627 | 24.7 | 25.5 | 5.661092e-07 | 1.657346e+06 |
| Patient/Treatment | 1.039216 | 26.5 | 26.5 | 5.576050e-12 | 1.936800e+11 |

* The median diagnostic interval duration is approximately 93% of the median patient interval duration.
* The median diagnostic interval duration is approximately 97% of the median treatment interval duration.
* The median patient interval duration is approximately 104% of the median treatment interval duration.

1. Intervals - Compare the durations of different intervals related to breast cancer across different income groups (high vs. lower income)

| **Interval Type** | **Number of Studies** | **Total N** | **Pooled Median** | **95% CI** | **Mann-Whitney U Test (High vs Lower Income)** |
| --- | --- | --- | --- | --- | --- |
| All Breast Cancer | | | | | |
| Patient Interval | 93 | 24,957.0 | 42.70 | [7.00, 382.08] | Statistic=518.0, p-value=0.0025\* |
| Diagnostic Interval | 56 | 95,053.0 | 24.50 | [3.13, 148.13] | Statistic=105.0, p-value=0.3392 |
| Treatment Interval | 99 | 1,709,349.0 | 29.00 | [9.25, 66.30] | Statistic=70.5, p-value=0.3531 |
| Non-Pregnancy Associated BC | | | | | |
| Patient Interval | 87 | 24,812.0 | 42.14 | [7.00, 386.04] | Statistic=403.5, p-value=0.0024\* |
| Diagnostic Interval | 51 | 94,808.0 | 25.00 | [6.00, 144.00] | Statistic=62.0, p-value=0.2938 |
| Treatment Interval | 95 | 1,709,116.0 | 30.00 | [8.75, 66.90] | Statistic=36.5, p-value=0.1372 |
| Pregnancy Associated BC | | | | | |
| Patient Interval | 6 | 145.0 | 61.00 | [30.63, 260.38] | Statistic=0.0, p-value=0.1002 |
| Diagnostic Interval | 5 | 245.0 | 8.00 | [1.10, 191.70] | Statistic=0.0, p-value=0.2765 |
| Treatment Interval | 4 | 233.0 | 20.00 | [19.00, 23.78] | Statistic=0.0, p-value=0.2765 |

* The median duration of the patient interval for all breast cancer patients is approximately 42.70 days, with a wide 95% confidence interval ranging from 7.00 to 382.08 days (p<0.05).
* The median duration of the diagnostic interval for all breast cancer patients is approximately 24.50 days, with a 95% confidence interval ranging from 3.13 to 148.13 days (p= 0.3392).
* The median duration of the treatment interval for all breast cancer patients is approximately 29.00 days, with a 95% confidence interval ranging from 9.25 to 66.30 days (p = 0.3531).

A graph of different colored squares

Description automatically generated

1. Intervals of High-income country

| **Cancer Type** | **Interval** | **Number of Studies** | **Total N** | **Pooled Median** | **95% CI Lower Bound** | **95% CI Upper Bound** |
| --- | --- | --- | --- | --- | --- | --- |
| All Breast Cancer | Patient | 26 | 6911.0 | 31.94 | 6.63 | 99.69 |
| All Breast Cancer | Diagnostic | 30 | 87053.0 | 19.0 | 1.73 | 46.77 |
| All Breast Cancer | Treatment | 70 | 1698176.0 | 29.5 | 10.63 | 47.65 |
| Non-Pregnancy Associated Breast Cancer | Patient | 22 | 6798.0 | 28.5 | 6.53 | 102.94 |
| Non-Pregnancy Associated Breast Cancer | Diagnostic | 26 | 86820.0 | 21.0 | 5.63 | 48.88 |
| Non-Pregnancy Associated Breast Cancer | Treatment | 66 | 1697943.0 | 30.85 | 10.13 | 48.25 |
| Pregnancy Associated Breast Cancer | Patient | 4 | 113.0 | 48.0 | 30.38 | 61.0 |
| Pregnancy Associated Breast Cancer | Diagnostic | 4 | 233.0 | 5.0 | 1.08 | 8.93 |
| Pregnancy Associated Breast Cancer | Treatment | 4 | 233.0 | 20.0 | 19.0 | 23.78 |

1. Patient Interval - All Breast Cancer:

The pooled median patient interval for all breast cancer patients is 31.94 days (95% CI: 6.63 - 99.69). This interval duration is significantly different from that of patients with pregnancy-associated breast cancer (p-value < 0.05), suggesting that patient intervals vary significantly between these two income groups.

1. Intervals of Low-income country

| **Cancer Type** | **Interval** | **Number of Studies** | **Total N** | **Pooled Median** | **95% CI Lower Bound** | **95% CI Upper Bound** |
| --- | --- | --- | --- | --- | --- | --- |
| All Breast Cancer | Patient | 67 | 18046.0 | 58.0 | 9.0 | 393.04 |
| All Breast Cancer | Diagnostic | 26 | 8000.0 | 55.0 | 11.63 | 173.25 |
| All Breast Cancer | Treatment | 29 | 11173.0 | 28.0 | 10.97 | 102.3 |
| Non-Pregnancy Associated Breast Cancer | Patient | 65 | 18014.0 | 50.0 | 9.0 | 393.48 |
| Non-Pregnancy Associated Breast Cancer | Diagnostic | 25 | 7988.0 | 53.0 | 11.4 | 147.0 |
| Non-Pregnancy Associated Breast Cancer | Treatment | 29 | 11173.0 | 28.0 | 10.97 | 102.3 |
| Pregnancy Associated Breast Cancer | Patient | 2 | 32.0 | 195.5 | 113.33 | 277.68 |
| Pregnancy Associated Breast Cancer | Diagnostic | 1 | 12.0 | 212.0 | 212.0 | 212.0 |

1. Patient Interval - All Breast Cancer:

The pooled median patient interval for all breast cancer patients from the low-income group is 58.0 days (95% CI: 9.0 - 393.04).

1. Diagnostic Interval - Non-Pregnancy Associated Breast Cancer:

The pooled median diagnostic interval for non-pregnancy associated breast cancer patients from the low-income group is 53.0 days (95% CI: 11.4 - 147.0).

1. Treatment Interval - All Breast Cancer:

The pooled median treatment interval for all breast cancer patients from the low-income group is 28.0 days (95% CI: 10.97 - 102.3).

Conclusion

Patient intervals in lower-income countries were consistently 0.5 to 4 times longer than those in high-income countries. Specifically, Intervals in high-income countries for all breast cancer patient interval 31.94 days, diagnostic interval 19.0 days, treatment interval 29.5 days compared to in low-income country patient interval 58.0 days, diagnostic interval 55.0 days, treatment interval 28.0 days. And for pregnancy associated breast cancer patient interval 195.5 days, diagnostic interval 212.0 days compare to patient interval 48.0 days, diagnostic interval 5.0 days, treatment interval 20.0 days in high-income countries. These findings highlighted global disparities in timely diagnosis and treatment of cancer, emphasizing the need to reduce help-seeking times for cancer symptoms in lower-income countries.

Models Selection:

1. Ridge and Lasso regression

Ridge and Lasso Regression are considered for this analysis despite the normality and linearity assumptions not being fully met. These techniques are chosen for their ability to prevent overfitting, as well as their simplicity, interpretability, and effectiveness in handling small observation datasets.

1. Huber Regression

Huber Regression, a robust regression technique, is selected to handle potential outliers or influential data points that may be present due to the varying health disparities among breast cancer patients across different income levels or countries. This method is less sensitive to outliers compared to ordinary least squares (OLS) regression, providing more reliable estimates by downweighing the impact of outliers while still maintaining interpretable coefficients.

Trade-offs: this model assumes that the majority of the data follows a normal distribution.

1. Linear Regression with Log Transformation

Linear Regression with Log Transformation is used to address the issue of non-normality in the residuals or target variable. By applying a logarithmic transformation to the target variable (e.g., patient interval, diagnostic interval, treatment interval), the resulting residual distribution can become more normally distributed. This transformation helps meet the assumptions of linear regression and improves the model's performance.

Trade-offs: Log transformation can help stabilize variance and improve the linearity of the relationship between predictors and the target.

However, it assumes that the relationship between predictors and the target is multiplicative rather than additive, which might not always be the case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficients | R-squared | MSE | MAE |
| Huber Regression | -31.78979443 | -0.24 | 25323.08 | 106.87 |
| Linear Regression with Log Transformation | -0.75650387 | -0.27 | 26009.31 | 108.15 |
| Ridge Regression | -61.22246331 | -0.11 | 22730.68 | 106.15 |
| Lasso Regression | -48.02897436 | -0.11 | 22733.79 | 106.12 |

* Huber Regression for Patient Interval: the coefficient of -31.7898 suggest that on average, the patient interval duration is shorter by approximately 32 days in high-income countries compared to lower-income countries. The R-squared value of -0.24 indicates that the model does not fit the data well.
* Linear Regression with Log Transformation: The R-squared value of -0.27 indicates that the model does not fit the data well, and the MSE (26009.31) and MAE (108.15) values are slightly higher than those of the Huber Regression, indicating that this model may have a higher prediction error.
* Ridge Regression for Patient Interval: the coefficient for the high-income country is -61.2225, which suggests that, on average, the patient interval duration is shorter by approximately 61 days in high-income countries compared to lower-income countries, after accounting for regularization. However, the negative value (-0.11) suggests that the models are performing worse than a simple baseline model that predicts the mean of the dependent variable.
* Lasso Regression for Diagnostic Interval: The coefficient is -48.0290, which suggests that the diagnostic interval duration is shorter by approximately 48 days in high-income countries compared to lower-income countries.

Discussion

I spent a significant amount of time converting the original R code to Python code, but I overlooked the nature of the dataset, particularly the fact that the data containing breast cancer is not significant. I also ignored missing outcome variables such as healthcare appointment accessibility, death from breast cancer, or other factors that may be more feasible in establishing a connection between delays in diagnosis or follow-up visits and health outcomes. This information would be more valuable for building an accurate model, such as adding more confounders or interpreting them as predictor variables instead of a single predictor.

Possible reasons for the poor performance of the models:

1. Lack of relevant features: The models may be missing important predictor variables that are strongly correlated with the dependent variable. With only one predictor (GNIG\_high), the models may not have enough information to make accurate predictions.
2. Non-linear relationships: If the relationship between the predictor and the dependent variable is non-linear, linear regression models (including Ridge and Lasso) may not be able to capture the underlying patterns effectively.
3. Data quality: After filtering the dataset to include only breast cancer patients, there are few observations remaining.

Records with 'all breast in high-income': 100

Records with 'all breast in low-income': 84

Number of all breast patient that both have PT, DI, TI value: 16

Number of PA breast patient that both have PT, DI, TI value: 4

Records with 'PA breast in high-income': 4

Records with 'PA breast in low-income': 0

Reporting of patient intervals was more common in studies from lower-income countries, while reporting of treatment intervals was more common in studies from high-income countries.

1. The study categorized countries into high- and lower-income economies based on indicators like gross national income (GNI) and Human Development Index (HDI). Other groups like the upper middle and lower middle were not considering subintervals in this dataset.

By addressing these limitations and incorporating additional relevant factors, future studies can develop more accurate and comprehensive models to address health disparities in breast cancer diagnosis and treatment across various income levels.