Medicare Physician Billing Patterns in RI and NH: A Comparative Analysis (2022–2023)

Shan Aziz

2025-06-31

Introduction

This report explores Medicare billing patterns among physicians practicing in Rhode Island (RI) and New Hampshire (NH) using the Medicare Physician & Other Practitioners by Provider dataset from the Centers for Medicare & Medicaid Services (CMS). We focus on providers with data available for both 2022 and 2023, allowing for a panel analysis at the provider-year level.

To ensure consistency and relevance, we restrict our analysis to physicians holding either a **Doctor of Medicine (MD)** or **Doctor of Osteopathic Medicine (DO)** degree, and whose practice location is within Rhode Island or New Hampshire. This allows for meaningful comparisons across states, specialties, and years.

The key objectives of this exercise are to:

- Summarize and compare **total submitted charges** and **total allowed charges** across the two states and years
- Identify the **three most common physician specialties** and their distribution within each state
- Estimate a **linear regression model** to test whether physicians in New Hampshire differ significantly in allowed charges compared to those in Rhode Island, controlling for specialty and year
- Calculate the **correlation** between submitted charges in 2022 and allowed charges in 2023 to assess year-to-year billing consistency

The goal of this study is to uncover state-level and specialty-level patterns in Medicare billing and evaluate the stability of physician payment behavior across years.

```
# Loading libraries
library(dplyr)
library(readr)
library(ggplot2)
library(knitr)
library(tidyr)
library(patchwork)
library(scales)
library(stringr)
library(broom)
```

Preliminary set up

```
# Loading both datasets
df_2022 <- read_csv("Medicare_Physician_Other_Practitioners_by_Provider_2022.csv")
df_2023 <- read_csv("Medicare_Physician_Other_Practitioners_by_Provider_2023.csv")

# Add year variable to each dataset before combining
df_2022$year <- 2022
df_2023$year <- 2023

# Combine both years into a single data frame
combined_df <- rbind(df_2022, df_2023)
```

Data source: https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider

The unit of observation should be provider-year, with two observations per provider (2022 and 2023)

```
# Get detailed breakdown of provider appearances
provider_counts <- table(combined_df$Rndrng_NPI)
provider_counts_df <- data.frame(
  Rndrng_NPI = names(provider_counts),
  count = as.numeric(provider_counts)
)

# See the distribution
table(provider_counts_df$count)
```

```
##
##       1       2
##  219812 1134912
```

```
# Find providers in both years (balanced panel)
balanced_providers <- provider_counts_df$Rndrng_NPI[provider_counts_df$count == 2]
cat("Balanced panel providers:", length(balanced_providers), "\n")
```

```
## Balanced panel providers: 1134912
```

```
# Find providers in only one year
single_year_providers <- provider_counts_df$Rndrng_NPI[provider_counts_df$count == 1]
cat("Single-year providers:", length(single_year_providers), "\n")
```

```
## Single-year providers: 219812
```

```r
# Keep only providers appearing in both years
balanced_df <- combined_df[combined_df$Rndrng_NPI %in% balanced_providers, ]

# Verify: should have exactly 2 observations per provider
table(table(balanced_df$Rndrng_NPI))
```

```
##
##       2
## 1134912
```

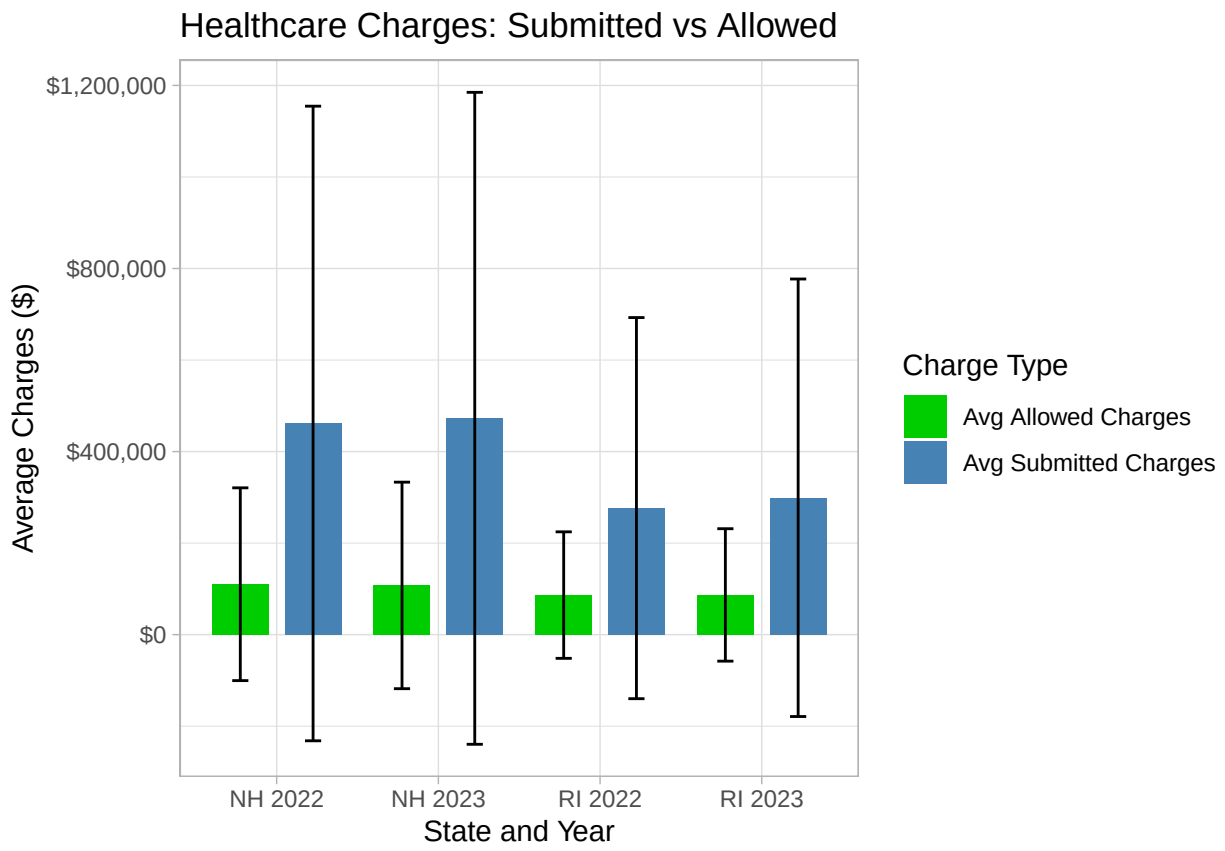
Filter for MD or DO and state is RI or NH

```r
# Advance cleaning step because of inconsistant MD and Do

md_do_df <- balanced_df %>%
  filter(
    Rndrng_Prvdr_Crdntls %>%
      str_replace_all("M[,\\s]?D", "MD") %>%       # fix "M,D", "M D", etc.
      str_replace_all("D[,\\s]?O", "DO") %>%       # fix "D,O", "D O", etc.
      str_to_upper() %>%                           #uppercase
      str_replace_all("\\.", "") %>%               #Removes all periods
      str_replace_all("\\s+", " ") %>%             #Replaces multiple spaces with a single space
      str_trim() %>%                               #Removes leading and trailing spaces
      str_detect("\\bMD\\b")                       #\\b = word boundary in regex
  )

filtered_df <- md_do_df %>%
  filter(Rndrng_Prvdr_State_Abrvtn %in% c("RI", "NH"))
```

1. A short table of summary statistics comparing the Medicare billings of Rhode Island and New Hampshire physicians. Comparing the annual mean and standard deviation of total submitted charges and total allowed charges treated by physician state.

```r
# Summary Statistics
summary_stats <- filtered_df %>%
  group_by(State = Rndrng_Prvdr_State_Abrvtn, Year=year) %>%
  summarise(
    `Avg Submitted Charges` = mean(Tot_Sbmtd_Chrg, na.rm = TRUE),
    `SD Submitted Charges` = sd(Tot_Sbmtd_Chrg, na.rm = TRUE),
    `Avg Allowed Charges` = mean(Tot_Mdcr_Alowd_Amt, na.rm = TRUE),
    `SD Allowed Charges` = sd(Tot_Mdcr_Alowd_Amt, na.rm = TRUE),
  )

# Now I need to reshape this data to make a comparison chart
# Want to plot submitted vs allowed side by side
plot_data <- summary_stats %>%
  select(State, Year, `Avg Submitted Charges`, `Avg Allowed Charges`) %>%
  pivot_longer(cols = c(`Avg Submitted Charges`, `Avg Allowed Charges`),
               names_to = "Charge_Type",
```

```
              values_to = "Amount")

# Reshape SDs
sd_data <- summary_stats %>%
  select(State, Year, `SD Submitted Charges`, `SD Allowed Charges`) %>%
  pivot_longer(cols = c(`SD Submitted Charges`, `SD Allowed Charges`),
               names_to = "Charge_Type",
               values_to = "SD") %>%
  mutate(Charge_Type = gsub("SD", "Avg", Charge_Type))  # Match names to join

# Join average and SD data
plot_data <- left_join(plot_data, sd_data,
                       by = c("State", "Year", "Charge_Type"))

# Create a combined label for the x-axis to show state and year together
plot_data$State_Year <- paste(plot_data$State, plot_data$Year)

# Make the comparison chart
chart <- ggplot(plot_data, aes(x = State_Year, y = Amount, fill = Charge_Type)) +
  geom_col(position = position_dodge(0.9), width = 0.7) +
  geom_errorbar(aes(ymin = Amount - SD, ymax = Amount + SD),
                position = position_dodge(0.9),
                width = 0.2, color = "black") +
  scale_fill_manual(values = c("Avg Submitted Charges" = "steelblue",
                               "Avg Allowed Charges" = "green3")) +
  scale_y_continuous(labels = dollar_format()) +
  labs(title = "Healthcare Charges: Submitted vs Allowed",
       x = "State and Year",
       y = "Average Charges ($)",
       fill = "Charge Type") +
  theme_light()

# Show the chart
print(chart)
```




```
# Also show the detailed numbers in a table
kable(summary_stats, caption = "Summary Statistics by State and Year")
```

Table 1: Summary Statistics by State and Year

State	Year	Avg Submitted Charges	SD Submitted Charges	Avg Allowed Charges	SD Allowed Charges
NH	2022	461374.8	693319.5	110174.03	210610.9
NH	2023	472662.3	712272.2	107526.93	225650.6
RI	2022	276441.8	416372.3	86406.25	138214.4
RI	2023	299083.3	477989.1	86783.69	144684.2

Key Insights:

- NH physicians submitted 50–70% more than RI physicians in both 2022 and 2023.
- Despite higher submissions, allowed charges only showed a 2.4% decrease in NH and 0.44% increase in RI from 2022 to 2023
- NH showed larger standard deviations, indicating more variation in physician billing behavior.

Three most common specialties in this data and reporting the proportion of doctors with these specializations within each state.

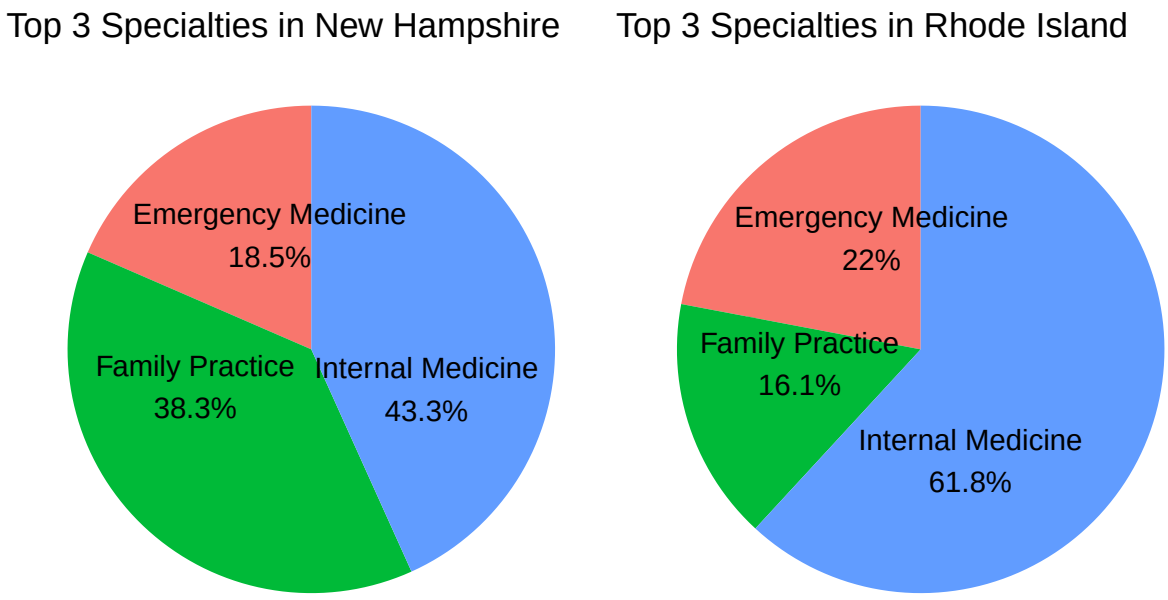
```
# Find the top 3 most common medical specialties in overall data
top_specialties <- filtered_df %>%
  count(Rndrng_Prvdr_Type) %>%
  arrange(desc(n)) %>%
  head(3)

# Extract just the specialty names for filtering
top_3_names <- top_specialties$Rndrng_Prvdr_Type

# Filter my data to only include these top 3 specialties
data_top3 <- filtered_df %>%
  filter(Rndrng_Prvdr_Type %in% top_3_names)

# Now calculate what percentage each specialty makes up within each state
specialty_summary <- data_top3 %>%
  group_by(State = Rndrng_Prvdr_State_Abrvtn, Specialty = Rndrng_Prvdr_Type) %>%
  count() %>%
  group_by(State) %>%
  mutate(Percentage = round(n / sum(n) * 100, 1))

# Show the detailed breakdown in a table
kable(specialty_summary, caption = "Proportion of Top 3 Specialties by State")
```

Table 2: Proportion of Top 3 Specialties by State

State	Specialty	n	Percentage
NH	Emergency Medicine	303	18.5
NH	Family Practice	628	38.3
NH	Internal Medicine	710	43.3
RI	Emergency Medicine	352	22.0
RI	Family Practice	258	16.1
RI	Internal Medicine	988	61.8

```
# Now lets visualize this with pie charts for each state
# function to make consistent pie charts

make_pie_chart <- function(state_data, state_name) {
  ggplot(state_data, aes(x = "", y = Percentage, fill = Specialty)) +
    geom_col(width = 1) +
    coord_polar(theta = "y") +
    geom_text(aes(label = paste0(Specialty, "\n", Percentage, "%")),
              position = position_stack(vjust = 0.7)) +
    labs(title = paste("Top 3 Specialties in", state_name)) +
    theme_void() +
    theme(legend.position = "none")
}

# Get data for each state separately
```

```r
nh_data <- specialty_summary %>% filter(State == "NH")
ri_data <- specialty_summary %>% filter(State == "RI")

# pie charts for both states
nh_pie <- make_pie_chart(nh_data, "New Hampshire")
ri_pie <- make_pie_chart(ri_data, "Rhode Island")

# both charts side by side for easy comparison (special thanks to patchwork lib)
nh_pie + ri_pie
```



2. Regression that tests whether physicians who are in New Hampshire (vs. Rhode Island) have higher total allowed charges, controlling for provider type (i.e., specialty) and year.

```r
# linear regression model
reg_model <- lm(Tot_Mdcr_Alowd_Amt ~ Rndrng_Prvdr_State_Abrvtn + Rndrng_Prvdr_Type + year, data = filte

# Tidy the model output
tidy_output <- tidy(reg_model)

# Accessing the 'term' column directly and use gsub for replacement
# Remove "Rndrng_Prvdr_Type"
tidy_output$term <- gsub("Rndrng_Prvdr_Type", "", tidy_output$term)
```

```r
# Remove "Rndrng_Prvdr_State_Abrvtn"
tidy_output$term <- gsub("Rndrng_Prvdr_State_Abrvtn", "", tidy_output$term)

#output
kable(tidy_output,
      digits = 4,
      format = "latex",    # Explicitly set format for PDF output
      longtable = TRUE)    # multi-page tables
```

term	estimate	std.error	statistic	p.value
(Intercept)	3278112.283	6966575.028	0.4705	0.6380
RI	-19082.562	3543.961	-5.3845	0.0000
Advanced Heart Failure and Transplant Cardiology	107108.176	120049.619	0.8922	0.3723
Allergy/ Immunology	36271.885	88573.273	0.4095	0.6822
Anesthesiology	30245.018	85212.858	0.3549	0.7226
Cardiac Surgery	185956.243	94901.353	1.9595	0.0501
Cardiology	133371.381	85383.963	1.5620	0.1183
Clinical Cardiac Electrophysiology	130890.880	93838.739	1.3948	0.1631
Colorectal Surgery (Proctology)	72620.397	91417.380	0.7944	0.4270
Critical Care (Intensivists)	55290.910	88344.561	0.6259	0.5314
Dermatology	267077.411	85754.319	3.1144	0.0018
Diagnostic Radiology	136338.512	85191.555	1.6004	0.1095
Emergency Medicine	39981.701	85137.532	0.4696	0.6386
Endocrinology	59004.201	86629.871	0.6811	0.4958
Family Practice	45039.708	85073.254	0.5294	0.5965
Gastroenterology	96370.817	85614.001	1.1256	0.2603
General Practice	110611.397	102011.815	1.0843	0.2783
General Surgery	40646.531	85336.412	0.4763	0.6339
Geriatric Medicine	59583.576	88345.304	0.6744	0.5000
Geriatric Psychiatry	87869.004	93842.044	0.9364	0.3491
Gynecological Oncology	51808.090	91176.845	0.5682	0.5699
Hand Surgery	151917.934	91963.762	1.6519	0.0986
Hematology	55210.998	96251.998	0.5736	0.5662
Hematology-Oncology	220178.800	86065.859	2.5583	0.0105
Hospice and Palliative Care	18253.855	88419.675	0.2064	0.8364
Hospitalist	49060.751	85645.684	0.5728	0.5668
Infectious Disease	25499.012	86526.969	0.2947	0.7682
Internal Medicine	50865.620	84979.063	0.5986	0.5495
Interventional Cardiology	110605.394	89126.589	1.2410	0.2146
Interventional Pain Management	249447.367	109593.746	2.2761	0.0229
Interventional Radiology	118360.743	90189.194	1.3124	0.1894
Medical Genetics and Genomics	-18706.989	120049.619	-0.1558	0.8762
Medical Oncology	326825.305	88283.201	3.7020	0.0002
Micrographic Dermatologic Surgery	542258.598	103954.693	5.2163	0.0000
Nephrology	149494.528	86436.344	1.7295	0.0837
Neurology	66467.081	85505.277	0.7773	0.4370
Neurosurgery	85445.257	86924.225	0.9830	0.3256
Obstetrics & Gynecology	4633.791	85294.796	0.0543	0.9567
Ophthalmology	433286.726	85560.453	5.0641	0.0000
Optometry	131267.021	147024.816	0.8928	0.3720
Oral Surgery (Dentist only)	9996.401	147024.816	0.0680	0.9458

Orthopedic Surgery	145962.827	85262.585	1.7119	0.0869
Osteopathic Manipulative Medicine	105795.954	147024.816	0.7196	0.4718
Otolaryngology	87704.324	86422.713	1.0148	0.3102
Pain Management	163789.249	90545.166	1.8089	0.0705
Pathology	55307.536	85740.481	0.6451	0.5189
Pediatric Medicine	28204.159	91965.310	0.3067	0.7591
Physical Medicine and Rehabilitation	85561.593	87307.425	0.9800	0.3271
Plastic and Reconstructive Surgery	39430.085	87617.768	0.4500	0.6527
Psychiatry	39162.571	85215.546	0.4596	0.6458
Pulmonary Disease	60614.073	85782.449	0.7066	0.4798
Radiation Oncology	241411.172	86925.288	2.7772	0.0055
Rheumatology	202620.777	86705.661	2.3369	0.0195
Sleep Medicine	51644.199	95531.414	0.5406	0.5888
Sports Medicine	93772.229	92274.088	1.0162	0.3095
Surgical Oncology	30658.372	93842.249	0.3267	0.7439
Thoracic Surgery	118249.699	89354.637	1.3234	0.1857
Undefined Physician type	132989.430	189810.603	0.7006	0.4835
Undersea and Hyperbaric Medicine	35785.814	147024.816	0.2434	0.8077
Urology	119185.933	85920.955	1.3872	0.1654
Vascular Surgery	118971.572	86789.303	1.3708	0.1705
year	-1610.218	3444.276	-0.4675	0.6401

Values of p	Inference
$p > 0.10$	No evidence against the null hypothesis.
$0.05 < p < 0.10$	Weak evidence against the null hypothesis
$0.01 < p < 0.05$	Moderate evidence against the null hypothesis
$0.05 < p < 0.001$	Good evidence against null hypothesis.
$0.001 < p < 0.01$	Strong evidence against the null hypothesis
$p < 0.001$	Very strong evidence against the null hypothesis

Figure 1: *https://itfeature.com/hypothesis/p-value-definition/*

Interpretation of results

We found a statistically significant difference in total allowed Medicare charges for physicians practicing in Rhode Island compared to the reference state. Specifically, Rhode Island providers had lower total allowed charges on average (Estimate = -19,083; p = 7.43e-08).

Specialty Effect

Several medical specialties were associated with significantly higher total allowed charges:

- Medical Oncology (p = 0.000215)
- Micrographic Dermatologic Surgery (p = 1.86e-07)
- Dermatology (p = 0.001848)

- Radiation Oncology ($p = 0.005493$)
- Hematology-Oncology ($p = 0.010535$)
- Interventional Pain Management ($p = 0.022861$)
- Rheumatology ($p = 0.019466$)

Other specialties showed marginal significance ($0.05 < p < 0.10$), including:

- Cardiac Surgery ($p = 0.0501$)
- Hand Surgery ($p = 0.0986$)
- Nephrology ($p = 0.0837$)
- Orthopedic Surgery ($p = 0.0869$)
- Pain Management ($p = 0.0705$)

These results suggest that specialty type is an important predictor of Medicare allowed charges.

Year Effect

There was no statistically significant linear trend in total allowed charges over time ($p = 0.6401$), indicating that Medicare payments did not meaningfully increase or decrease across the years analyzed.

Model Fit

The model explains approximately 19% of the variation in total allowed charges (Adjusted $R^2 = 0.1865$). This indicates that while the included variables contribute meaningfully, a large portion of the variation remains unexplained, suggesting the presence of other influential factors not captured in the model.

3. Based on the regression above, Specialty with the highest allowed charges (controlling for state and year)?

```
# I want to find which medical specialty gets paid the most
# plan: extract coefficients -> filter for specialties -> find the highest one

# Get all coefficients from regression model
model_results <- coef(reg_model)

# I only care about specialty coefficients, not state/year effects
# These all have "Rndrng_Prvdr_Type" in their names
specialty_results <- model_results[grepl("Rndrng_Prvdr_Type", names(model_results))]

# This line looks at the names of the 'specialty_results'
# and replaces "Rndrng_Prvdr_Type" with an empty space ("").
names(specialty_results) <- gsub("Rndrng_Prvdr_Type", "", names(specialty_results))

# Sort them highest to lowest and grab the top one
highest_allowed_charges <- sort(specialty_results, decreasing = TRUE) %>% head(1)

# Printing highest allowed charges
print(highest_allowed_charges)
```

```
## Micrographic Dermatologic Surgery
##                           542258.6
```

Micrographic Dermatologic Surgery had the highest average total allowed charges, with an estimated value of $542,258.59

4. Calculating the correlation between physicians' total submitted charges in 2022 and total allowed charges in 2023.

```r
# Create 2022 dataset with submitted charges
charges_2022 <- filtered_df %>%
  filter(year == 2022) %>%
  select(Rndrng_NPI, Tot_Sbmtd_Chrg) %>%
  rename(submitted_2022 = Tot_Sbmtd_Chrg)

# Create 2023 dataset with allowed charges
charges_2023 <- filtered_df %>%
  filter(year == 2023) %>%
  select(Rndrng_NPI, Tot_Mdcr_Alowd_Amt) %>%
  rename(allowed_2023 = Tot_Mdcr_Alowd_Amt)

# Only keep physicians who appear in both years for fair comparison
cor_df <- inner_join(charges_2022, charges_2023, by = "Rndrng_NPI")

# Calculate the correlation between 2022 submissions and 2023 allowances
correlation_result <- cor(cor_df$submitted_2022, cor_df$allowed_2023, use = "complete.obs")

# Show the result
print(paste("Correlation (r) =", round(correlation_result, 3)))
```

```
## [1] "Correlation (r) = 0.696"
```

RANGE	INTERPRETATION
0<r<0.2	no or negligible correlation
0.2<r<0.4	low degree of correlation
0.4<r<0.6	moderate degree of
0.6<r<0.8	marked degree of
0.8<r<1	high correlation

Figure 2: *https://financetrainingcourse.com/education/2011/04/correlation-correlation-coefficient-r/*

Moderately strong positive correlation (r=0.659): Physicians who submitted higher charges in 2022 tend to have higher allowed charges in 2023, suggesting consistent billing patterns across years.

Helpful Resources Links:
https://stackoverflow.com/questions/61635604/r-pivot-longer-and-ggplot-errorbar-with-two-name-key-columns

https://www.youtube.com/watch?v=Oe5O4LRj2rc&t=1972s
https://stackoverflow.com/questions/61635604/r-pivot-longer-and-ggplot-errorbar-with-two-name-key-columns

https://www.geeksforgeeks.org/r-language/how-to-calculate-percentage-by-group-in-r/

https://r-graph-gallery.com/414-map-multiple-charts-in-ggiraph.html

https://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

https://stackoverflow.com/questions/30159162/linear-model-with-categorical-variables-in-r

https://www.statology.org/r-cor-function/

https://www.geeksforgeeks.org/r-language/compute-the-correlation-coefficient-value-between-two-vectors-in-r-programming-cor-function/