Final Project

Diyanet Nijiati

2025-05-28

## Motivation

Medical debt is a powerful marker of structural inequity in the U.S. healthcare system. Unlike other forms of debt, medical debt is rarely incurred by choice and often reflects structural deficiencies in insurance coverage, and public health infrastructure. Individuals from low-income and historically marginalized communities are disproportionately affected, facing heightened risks of financial hardship, stress, and delayed or forgone medical care.

Despite the scale and impact of debt relief efforts, the broader consequences of medical debt on healthcare utilization, mental health, and persistent disparities remain underexplored in policy and research. For example, does abolishing medical debt encourage people to return for needed care? Are there meaningful differences in how medical debt affects access to care and stress across demographic groups? And how should financial wellness be measured beyond conventional credit metrics?

My lab is partnering with Undue Medical Debt (UMD), a nonprofit that has abolished over $15 billion in medical debt for more than 10 million low-income and vulnerable Americans since 2014, to generate evidence on the broader impacts of debt relief beyond credit score improvements. In this project, I will leverage 2023 panel of Survey of Income and Program Participation (SIPP) data to examine the relationships between medical debt, healthcare utilization, and mental health. In specific, (1) does owing medical debt reduce an individual’s likelihood of seeking medical care? (2) how this effect varies by race, sex, and income, and (3) the association between debt burden and psychological distress. Clarifying these relationships will inform targeted policies and interventions to alleviate financial stress, improve access, and prevent avoidable health complications.

pu <- fread("pu2023.csv", sep = "|", select = c(  
 "SSUID","PNUM","MONTHCODE","ERESIDENCEID","ERELRPE","SPANEL", # IDs  
 "ESEX","TAGE","EEDUC","TRACE","TPTOTINC", # Demographics  
 "TMED\_AMT","EDEBT\_MED", #Medical debt  
 "TVISDOC", #Doctors visits  
 "RCONDTYP1" #Health conditions  
))  
  
# filter December 2023 data  
pu12 <- pu %>% filter(SPANEL == 2023, MONTHCODE == 12, TAGE >= 21)

…

pu12 <- pu12 %>% mutate(Sex = ifelse(ESEX == 1, 1, 0),  
 Age = as.numeric(TAGE),  
 DocVisit = as.numeric(TVISDOC),  
 VisitedDoc = ifelse(DocVisit > 0, 1, 0),  
 LogIncome = ifelse(TPTOTINC > 0, log(TPTOTINC + 1), NA),  
 HasDebt = ifelse(EDEBT\_MED == 1, 1, 0),  
 LogDebtAmount = ifelse(HasDebt == 1, log(TMED\_AMT + 1), NA),  
 MentalHealth = ifelse(RCONDTYP1 == 9, 1, 0))

## Data Description

I used 2023 panel of the Survey of Income and Program Participation (SIPP), administered by the U.S. Census Bureau. It follows a sample of civilian, non-institutionalized U.S. residents, collecting detailed information on income, program participation, health, and demographics. For this project, the analytic sample is restricted to adults age 21 and older, and records are limited to the December wave to ensure that annual summary variables, such as medical debt and doctor visits, are consistently measured.

#### Key Variables

Medical debt: HasDebt(binary indicator for whether any medical debt is owed, derived from EDEBT\_MED); LogDebtAmount(continuous, log transformed unpaid medical debt from TMED\_AMT).

Healthcare utilization: DocVisit(count of any medical provider visits in the past year, TVISDOC); VisitedDoc(binary indicator for whether the individual had any doctor visits).

Demographics: age(continuous), sex(binary), race(categorical), education(categorical), income (LogIncome, log transfromed from monthly income).

MentalHealth(binary, equals 1 if respondent reported a mental or emotional disorder RCONDTYP1 == 9).

#### Missing Data

# identify missing data  
miss\_var\_summary(pu12)

## # A tibble: 25 × 3  
## variable n\_miss pct\_miss  
## <chr> <int> <num>  
## 1 TMED\_AMT 8908 90.4   
## 2 LogDebtAmount 8908 90.4   
## 3 RCONDTYP1 5951 60.4   
## 4 MentalHealth 5951 60.4   
## 5 LogIncome 867 8.79  
## 6 SSUID 0 0   
## 7 PNUM 0 0   
## 8 MONTHCODE 0 0   
## 9 ERESIDENCEID 0 0   
## 10 ERELRPE 0 0   
## # ℹ 15 more rows

Medical debt amount and log transformed debt are about 90% missing. Only those with medical debt are asked about the amount. A small number of respondents reported having medical debt but a debt amount of zero. These cases were addressed within a two-part modeling approach: modeling debt presence for all, and modeling debt amount (log-transformed) only among those reporting any debt.

Health condition has about 60% missing. This question was only asked if the participant answered “yes” to at least one disability-related question. In analysis, this is only modeled for those who is observed, missing values are not imputed, so it could be a limitation for the analysis.

Some respondents reported negative or zero for monthly income, which are coded as missing income data. The missingness is not at random and may reflect some financial instability issues.

## Primary Model

mean(pu12$DocVisit)

## [1] 5.264631

var(pu12$DocVisit)

## [1] 115.0954

The mean number of doctor’s visits is 5.26, while the variance was much larger at 115.1.A standard Poisson regression might not fit the data due to overdispersion and excess zeros. So I chose to use Negative Binomial model. However, since nearly 20% of respondents reported no doctor visits, I also implemented two-part model for additional clarity and to more directly address my primary research question: (1) a logistic regression predicting the likelihood of any doctor visit, (2) a Negative Binomial regression predicting the frequency of visits among those with at least one visit.

# Do people with medical debt avoid seeing doctor entirely?  
logit\_model <- glm(VisitedDoc ~ HasDebt + Age + Sex + Race + Educ + LogIncome, data = pu12, family = binomial(link='logit'))  
summary(logit\_model)

##   
## Call:  
## glm(formula = VisitedDoc ~ HasDebt + Age + Sex + Race + Educ +   
## LogIncome, family = binomial(link = "logit"), data = pu12)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.443277 0.163981 -2.703 0.00687 \*\*   
## HasDebt 0.538524 0.103108 5.223 1.76e-07 \*\*\*  
## Age 0.029305 0.001574 18.622 < 2e-16 \*\*\*  
## Sex -0.677710 0.056229 -12.053 < 2e-16 \*\*\*  
## RaceAIAN -0.181553 0.231919 -0.783 0.43373   
## RaceAsian -0.154748 0.113976 -1.358 0.17455   
## RaceBlack -0.203274 0.081893 -2.482 0.01306 \*   
## RaceMultiracial 0.176755 0.193304 0.914 0.36051   
## RaceOther -0.395030 0.321296 -1.229 0.21889   
## EducBachelor 0.473842 0.074450 6.365 1.96e-10 \*\*\*  
## EducMasters or above 0.786937 0.097648 8.059 7.70e-16 \*\*\*  
## EducSome college or Associate 0.353515 0.068363 5.171 2.33e-07 \*\*\*  
## LogIncome 0.047343 0.018402 2.573 0.01009 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 9186.2 on 8991 degrees of freedom  
## Residual deviance: 8489.1 on 8979 degrees of freedom  
## (867 observations deleted due to missingness)  
## AIC: 8515.1  
##   
## Number of Fisher Scoring iterations: 4

exp(0.538524) # with medical debt

## [1] 1.713476

exp(0.029305) # age

## [1] 1.029739

exp(-0.677710) # sex

## [1] 0.5077785

exp(-0.203274) # black

## [1] 0.8160546

exp(0.786937) # masters or above

## [1] 2.196658

exp(0.047343) # income

## [1] 1.048482

People who have medical debt have about 71% higher odds of having any doctor visit, compared to people without debt, holding all else constant (and it is statistically significant, p < 0.01). This is somewhat different than the expected direction I wanted to go. I expected people with debt would visit doctor less, but the result also makes sense. It could be that people with debt often have ongoing medical issues that require frequent visits, leading both to increased medical debt and healthcare utilization.

Each additional year increases the odds by about 3%. This result also makes sense since older people are more likely to visit a doctor.

Compared to male respondents, females have about 49% lower odds of visiting at least once, the result is statistically significant (p < 0.01). Black individuals have statistically significant (p < 0.01) lower odds (about 18%) of having a visit compared to white individuals.

Individuals with a Masters or above have 2.196 times higher odds of having a doctor visit, indicating more than double the odds of visiting a doctor compared to those with only a high school diploma. This result is statistically significant (p < 0.01).

Higher income slightly increases the odds of having a visit, each one-unit increase in log income increases the odds of visiting a doctor about 5%, the result is statistically significant (p < 0.01).

# Among those who do seek care, does medical debt affect the frequency of visits?  
pu12\_doc <- pu12 %>% filter(DocVisit > 0)

## filter: removed 2,178 rows (22%), 7,681 rows remaining

model <- glm.nb(DocVisit ~ HasDebt + Age + Sex + Race + Educ + LogIncome, data = pu12\_doc)  
summary(model)

##   
## Call:  
## glm.nb(formula = DocVisit ~ HasDebt + Age + Sex + Race + Educ +   
## LogIncome, data = pu12\_doc, init.theta = 1.136116337, link = log)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.8015005 0.0799188 22.542 < 2e-16 \*\*\*  
## HasDebt 0.4013574 0.0391842 10.243 < 2e-16 \*\*\*  
## Age 0.0081360 0.0006909 11.777 < 2e-16 \*\*\*  
## Sex -0.1563791 0.0247698 -6.313 2.73e-10 \*\*\*  
## RaceAIAN 0.1202296 0.1158853 1.037 0.299508   
## RaceAsian -0.1948340 0.0555781 -3.506 0.000456 \*\*\*  
## RaceBlack -0.1165952 0.0388354 -3.002 0.002680 \*\*   
## RaceMultiracial 0.0243623 0.0882574 0.276 0.782520   
## RaceOther 0.1213900 0.1738558 0.698 0.485038   
## EducBachelor -0.0549932 0.0338453 -1.625 0.104197   
## EducMasters or above 0.0238175 0.0385083 0.619 0.536244   
## EducSome college or Associate 0.0370132 0.0311292 1.189 0.234433   
## LogIncome -0.0403238 0.0088814 -4.540 5.62e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(1.1361) family taken to be 1)  
##   
## Null deviance: 7679.4 on 7124 degrees of freedom  
## Residual deviance: 7318.2 on 7112 degrees of freedom  
## (556 observations deleted due to missingness)  
## AIC: 42075  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 1.1361   
## Std. Err.: 0.0202   
##   
## 2 x log-likelihood: -42046.6010

exp(0.4013574) # with medical debt

## [1] 1.493851 …

Among individuals who had at least one visit, those with medical debt have approximately 49.3% higher rate of doctor visits than those without medical debt. The result is statistically significant(p < 0.01). Again, suggests that individuals who owe medical debt might have ongoing, chronic, or severe health problems requiring more frequent care.

Older age positively predicts more visits. Each additional year of age is associated with 0.8% increase in the rate of doctor visit, which is statistically significant (p < 0.01). Females who have at least one visit typically have fewer visits compared to males, 14.5% lower rate of doctor visits.

Asian and Black respondents both have fewer visits compared to White respondents who have at least one visit. Asian individuals 18% lower rate, black individuals 11% lower. Other race categories are not statistically significant.

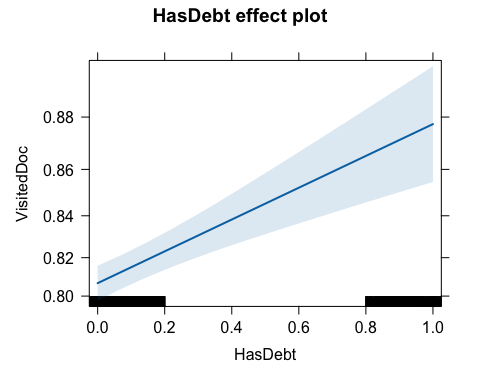
Education levels are not statistically significant among people who have at least one visit, suggesting education mainly influences decision to visit or not.

Higher income is associated with a 4% lower rate of doctor visits, which is different than what I found earlier.

## Graphic representation

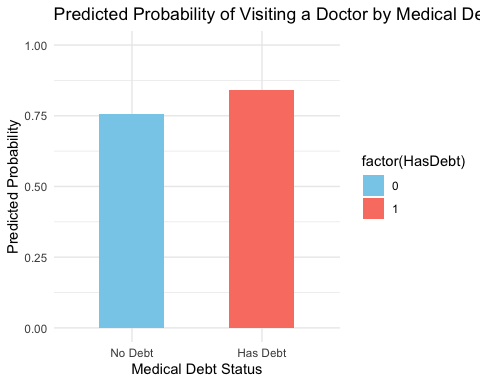
To visualize the overall relationship between medical debt and the probability of visiting a doctor, I first generate an effect plot from my logistic regression model. This plot shows how the predicted probability of visiting a doctor changes as medical debt status changes, while holding other variables constant.

plot(Effect(focal.predictors = "HasDebt", logit\_model))



To make the results even more accessible, I also made a barplot of the predicted probabilities for each group (with and without medical debt), holding all other variables at their mean or reference values. This approach provides a straightforward visual comparison of the likelihood of seeking care between the two groups, directly addressing the research question.

# create dataset for prediction  
pred\_data <- data.frame(  
 HasDebt = c(0, 1),  
 Age = mean(pu12$Age, na.rm = TRUE),  
 Sex = mean(pu12$Sex, na.rm = TRUE),  
 LogIncome = mean(pu12$LogIncome, na.rm = TRUE),  
 Race = "White",  
 Educ = "HS diploma or less"  
)  
  
pred\_data$Race <- factor(pred\_data$Race, levels = levels(pu12$Race))  
pred\_data$Educ <- factor(pred\_data$Educ, levels = levels(pu12$Educ))  
  
# predict probabilities  
pred\_data$pred\_prob <- predict(logit\_model, newdata = pred\_data, type = "response")  
  
ggplot(pred\_data, aes(x = factor(HasDebt, labels = c("No Debt", "Has Debt")), y = pred\_prob, fill = factor(HasDebt))) +  
 geom\_bar(stat = "identity", width = 0.5) +  
 scale\_fill\_manual(values = c("skyblue", "salmon")) +  
 labs(title = "Predicted Probability of Visiting a Doctor by Medical Debt",  
 x = "Medical Debt Status",  
 y = "Predicted Probability") +  
 theme\_minimal() +  
 ylim(0,1)



This graph shows the predicted probability of having at least one doctor visit in the past year, as estimated by the logistic regression model and holding other covariates at their mean or reference levels. Individuals with medical debt have a higher predicted probability of seeing a doctor compared to those without medical debt (about 75% vs. 85%). Similar to the regression result, this might suggest that people with medical debt are more likely to use healthcare, because those with ongoing health needs both accumulate debt and require frequent doctor visits, leading to a positive association between debt and utilization.

## Additional model

# interaction with income level  
# test whether debt effect differs significantly by income  
model\_income <- glm.nb(DocVisit ~ HasDebt \* LogIncome + Age + Sex + Educ + Race, data = pu12)  
summary(model\_income)

##   
## Call:  
## glm.nb(formula = DocVisit ~ HasDebt \* LogIncome + Age + Sex +   
## Educ + Race, data = pu12, init.theta = 0.6703663446, link = log)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.1672894 0.0919251 12.698 < 2e-16 \*\*\*  
## HasDebt 0.3703240 0.2851792 1.299 0.194093   
## LogIncome -0.0325556 0.0103014 -3.160 0.001576 \*\*   
## Age 0.0145417 0.0007802 18.638 < 2e-16 \*\*\*  
## Sex -0.3181906 0.0281425 -11.306 < 2e-16 \*\*\*  
## EducBachelor 0.0472191 0.0382684 1.234 0.217243   
## EducMasters or above 0.1772325 0.0446672 3.968 7.25e-05 \*\*\*  
## EducSome college or Associate 0.1124247 0.0352539 3.189 0.001428 \*\*   
## RaceAIAN 0.0497940 0.1282264 0.388 0.697773   
## RaceAsian -0.2132846 0.0622067 -3.429 0.000607 \*\*\*  
## RaceBlack -0.1478334 0.0435725 -3.393 0.000692 \*\*\*  
## RaceMultiracial 0.0711231 0.1001389 0.710 0.477552   
## RaceOther -0.0632498 0.1873009 -0.338 0.735596   
## HasDebt:LogIncome 0.0170642 0.0361328 0.472 0.636738   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.6704) family taken to be 1)  
##   
## Null deviance: 10607.5 on 8991 degrees of freedom  
## Residual deviance: 9954.7 on 8978 degrees of freedom  
## (867 observations deleted due to missingness)  
## AIC: 48231  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.6704   
## Std. Err.: 0.0116   
##   
## 2 x log-likelihood: -48200.7530

I tested whether the effect of medical debt on healthcare utilization (number of doctor visits) varies by income level by including an interaction between HasDebt and LogIncome. The interaction term was small and statistically insignificant (p = 0.64), indicating that the association between medical debt and doctor visits does not differ by income level in this sample.

model\_mental <- glm(MentalHealth ~ HasDebt + Age + Sex + Race + Educ + LogIncome, data = pu12, family = binomial)  
summary(model\_mental)

##   
## Call:  
## glm(formula = MentalHealth ~ HasDebt + Age + Sex + Race + Educ +   
## LogIncome, family = binomial, data = pu12)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.114630 0.290433 3.838 0.000124 \*\*\*  
## HasDebt -0.025954 0.133431 -0.195 0.845772   
…

To address the secondary question of whether medical debt is associated with psychological distress, I used a logistic regression model predicting whether respondents reported a mental or emotional disorder. The coefficient for medical debt was not statistically significant (p = 0.85), suggesting that in this sample, owing medical debt was not associated with increased odds of reporting a mental or emotional disorder, after adjusting for sociodemographic factors.

I wanted run another negative binomial regression among individuals with debt, but this time using lag-transformed medical debt amount as key predictor. I wonder if with the debt increases, the marginal impact of debt would decrease.

pu12\_debt <- pu12 %>% filter(HasDebt == 1, !is.na(LogDebtAmount), !is.na(DocVisit))

## filter: removed 8,908 rows (90%), 951 rows remaining

model <- glm.nb(DocVisit ~ LogDebtAmount + Age + Sex + Race + Educ + LogIncome, data = pu12\_debt)  
summary(model)

##   
## Call:  
## glm.nb(formula = DocVisit ~ LogDebtAmount + Age + Sex + Race +   
## Educ + LogIncome, data = pu12\_debt, init.theta = 0.7030817975,   
## link = log)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.182053 0.343006 3.446 0.000569 \*\*\*  
## LogDebtAmount 0.035724 0.017312 2.064 0.039062 \*   
…

exp(0.035724)

## [1] 1.03637

exp(0.035724 \* 2.3)

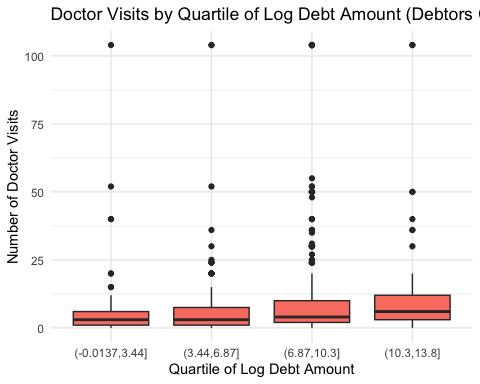
## [1] 1.085635

As we see from the coefficient, each one unit increase in log debt amount is associated with a 3.6% increase in the expected number of doctor visit, contolling for all other variables. And for context, log(1000+1) = 6.91, log(10000+1) = 9.21, the difference is about 2.3, so going from 1000 dollars to 10,000 dollars in medical debt is associated with about an 8.6% increase in visit frequency. This supports the hypothesis that the marginal effect of additional debt deminishes as the total owed increases, each extra dollar has less impact at higher debt level.

## Graphic representation

To explore how the amount of medical debt relates to healthcare use, I created a boxplot of the number of doctor visits by quartile of debt amount (for those with debt). I think the boxplot is an effective way to show the distribution, median, and spread of visit counts within each debt group, can highlight both typical experiences and the presence of outliers.

ggplot(pu12\_debt, aes(x = cut(LogDebtAmount, breaks=4), y = DocVisit)) +  
 geom\_boxplot(fill = "salmon") +  
 labs(title = "Doctor Visits by Quartile of Log Debt Amount (Debtors Only)",  
 x = "Quartile of Log Debt Amount",  
 y = "Number of Doctor Visits") +  
 theme\_minimal()



From the boxplot, we can see as debt amounts increase, there are more individuals with unusually high numbers of doctor visits, especially visible as outliers in the third quartile, suggesting that a subset of higher-debt individuals may be those with chronic or severe health issues requiring frequent care. This supports the regression finding earlier. However, the pattern is not strictly linear, and the highest debt group (fourth quartile) does not have more outliers than the lowest, I wonder if this is suggesting something here. One possibility is that people in the highest debt group may have already accumulated significant debt from prior healthcare utilization, and are now may be limiting additional care. But alternatively, the lower number of outliers in the top quartile could simply be a function of small sample sizes or random variation.

## Summary

This analysis provides some new evidence on the relationship between medical debt and healthcare utilization in the United States. Contrary to my original hypothesis, the results consistently show that individuals with medical debt are actually more likely to seek medical care than those without debt, both in terms of the probability of having any doctor visit and in the frequency of visits among those who do see a provider. This likely reflects that people who develop medical debt often do so precisely because they have higher and ongoing health needs, rather than because debt deters care outright.

Additional models found that the association between debt and doctor visits does not significantly differ by income, and there was no statistically significant link between medical debt and self-reported mental or emotional disorders after controlling for demographics. Among those with medical debt, having a higher debt amount is associated with more doctor visits, though the effect size is modest and the pattern is not strictly linear. Descriptive plots suggest that, at the highest debt levels, some individuals may be limiting further care, possibly due to financial exhaustion or barriers, though this pattern could also be due to sample size or random variation.

Overall, these findings highlight the complex relationship between financial barriers and healthcare utilization. While medical debt does not appear to prevent care for most people, it is closely linked to higher healthcare use, likely reflecting the chronic or acute health conditions that both drive and result from debt. These results suggest that policies focused solely on debt relief may not fully address the underlying health needs that lead to debt accumulation (I don’t know if I can tell Undue Medical Debt this, after all they are doing a good job helping people), and that ongoing monitoring and support for high-need populations is essential.