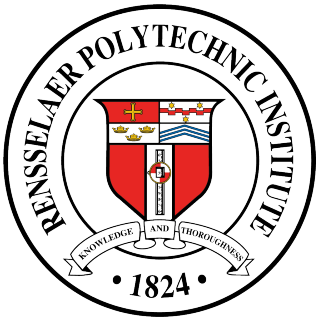
Insurance, Spending, & Claim Denials

An Analysis of Claim Denials

Max Troeger

MGMT 6600 - Data Analytics

Dr. Ahmed Eleish



Department of Economics

Rensselaer Polytechnic Institute

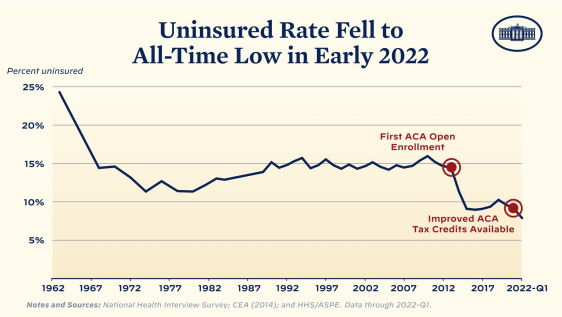
Troy, NY, USA

23 Apr. 2025

Contents

1 Abstract & Introduction

The Patient Protection and Affordable Care Act (ACA) of 2010 implemented federal health insurance subsidies and imposed an individual penalty, collected through tax return forms, for not carrying health insurance (Auerbach et al., 2010). As of 2023, 7.6% of Americans (roughly 1 in 13 people) do not carry any kind of health insurance (CDC, 2024). Paradoxically, many of these people are eligible for subisdies but do not take advantage of them (Baicker et al., 2012).



In 2018, the individual mandate penalty was repealed, and, starting in Q1 2022, federal subsidies available under the ACA were expanded. Despite the clear behavioral explanations for prevalent uninsurance in the United States, there may be practical reasons for uninsurance. In particular, Medical Expenditure Panel Survey (MEPS) survey data indicates that although costs bar access to medical care less and less, insurance issues in particular seem to be rising (ARHQ, 2025). In the light of the murder of UnitedHealthcare CEO Brian Thompson, Dyer (2025) writes that the American public generally views health insurance companies as greedy and unwilling to accept claims. The aim of this paper is to investigate if health insurance denials have been increasing systematically. We find that, far from increasing with time, health insurance claim denials seem to increase as a fixed proportion of health insurance claims received. Accounting for time and state fixed effects reveals that claim denials appear to track with the direction of insurance uptake.

2 Data Description

The Kaiser Family Foundation (KFF) provides robust datasets for health insurance information for health policy research and polling (Justin Lo and Wallace, 2025; Karen Pollitz and Wallace, 2023). In particular, KFF cleans up claim denials and appeals datasets produced by the Centers for Medicare and Medicaid Services (CMS) for non-group qualified health plans (QHPs) sold on HealthCare.gov.

We take 9 annual datasets covering 2015–2023 to generate panel data. Combining these datasets with `rbind` and unifying the headers gives the in and out-of-network approval and denial rates for each insurance provider on HealthCare.gov and organizes providers by their location and unique *issuer ID*. The dataset also provides annual claim denial, internal appeal, and disenrollment rates for each provider.

When combining these datasets we drop all variables save `State`, `Issuer_Name`, `Issuer_Id`, `Claims_Received`, `Claims_Denied`, and `Denial_Rate`. We create the variable `Year` to keep track of the year of each respective dataset.

Over the period of 2015–2023, we see that ACA QHPs denied of 17.8% of claims: insurance companies denied 425,327,006 of the 2,387,772,664 claims received. Figure 1 indicates that on the individual firm level, QHP providers denied a median of 17.2% of claims.

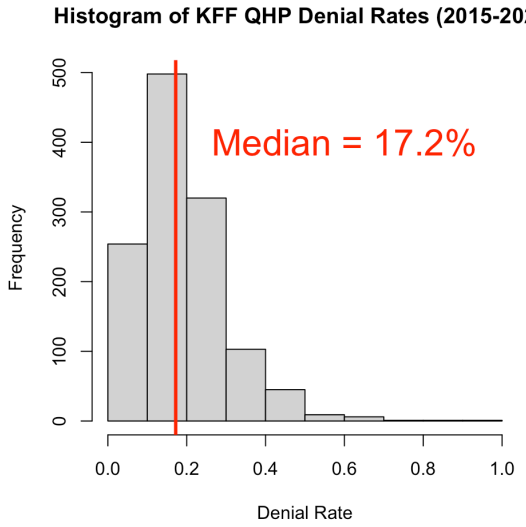


Figure 1: Exploratory analysis of aggregate denial rates

When we break the denial rate down into the number of overall claims and denials, we see skewed distributions for both data points. In the economics literature, we frequently transform variables with a base-10 logarithmic transformation to reduce the effect of heteroskedasticity and skewness.

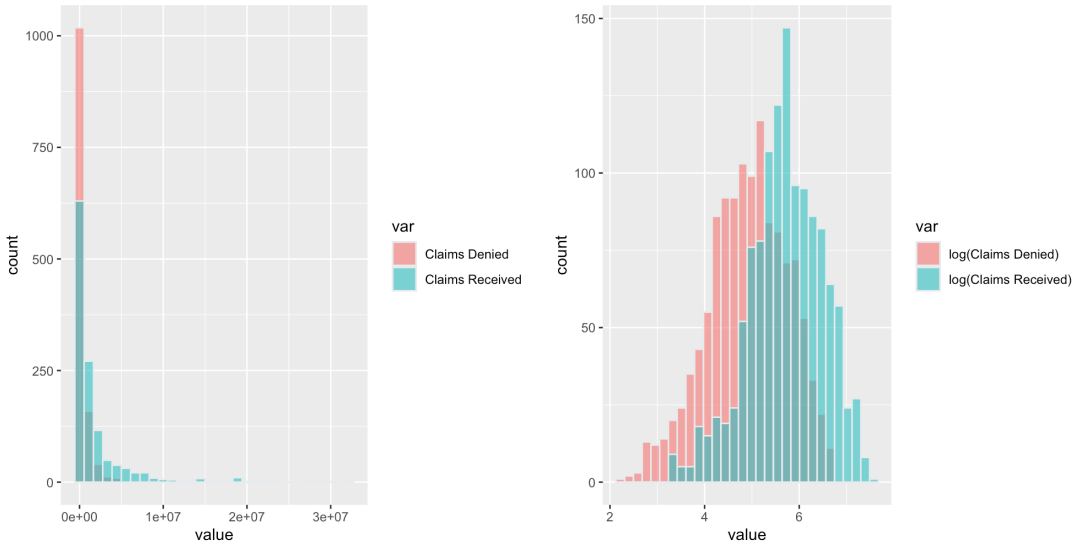


Figure 2: Exploratory analysis of aggregate denials and claims

Taking the logarithm of `Claims_Received` and `Claims_Denied` produces roughly normal distributions. Figure 2 illustrates the distributions of denials and claims before and after the logarithmic transformation.

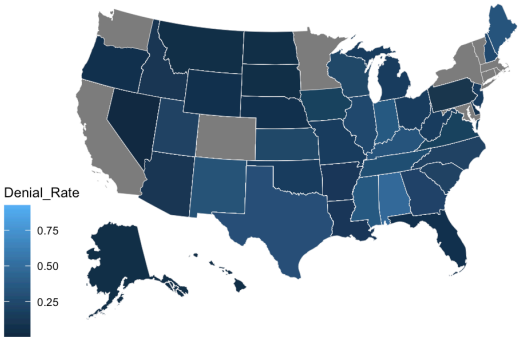


Figure 3: Denial rates by US state of provider origin

Unfortunately, the KFF datasets do not contain information for *state* insurance markets. As a consequence, Massachusetts, California, New York, and several other states are not included in the dataset. Figure 3 shows the intensity of health insurance claim denial rates by US state.

3 Exploratory Analysis

To explore the dataset we fit a model of the form

$$\widehat{Claims\_Denied} = \hat{\beta}_0 + \hat{\beta}_1 \widehat{Claims\_Received}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44857.8401	9450.0426	4.75	0.0000***
Claims_Received	0.1549	0.0023	68.42	0.0000***

Where $\bar{R}^2 = 0.791$ and both regressors are significant at the $\alpha = 0.1$[1]. We graph this regression visually in Figure 4.

[1]This confidence level is the norm in economics literature.

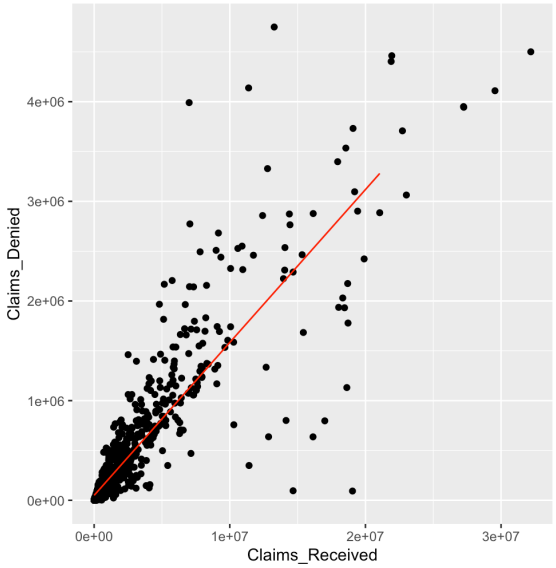


Figure 4: Claims denied vs. claims received

We now check to see if variance in the rate of denials can be explained by the number of claims received by the formula

$$\widehat{Denial\_Rate} = \hat{\beta}_0 + \hat{\beta}_1 \widehat{Claims\_Received}$$

Nevertheless, we get virtually meaningless results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1900	0.0041	46.44	0.0000
Claims_Received	-0.0000	0.0000	-1.31	0.1890

Where $\bar{R}^2 \approx 0$. Clearly a model of the former form produces stronger results.

We now consider time factors

$$\widehat{Claims\_Denied} = \hat{\beta}_0 + \hat{\beta}_1 \widehat{Claims\_Received} + (\text{time factors})$$

and receive the following model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50934.9762	27368.4933	1.86	0.0630*
Claims_Received	0.1547	0.0023	68.34	0.0000***
as.factor(Year)2016	-23091.1756	38648.6327	-0.60	0.5503
as.factor(Year)2017	5956.9077	37428.7030	0.16	0.8736
as.factor(Year)2018	-84860.9114	39934.0102	-2.13	0.0338**
as.factor(Year)2019	-14959.4755	38011.3673	-0.39	0.6940
as.factor(Year)2020	-5099.6237	36553.1325	-0.14	0.8891
as.factor(Year)2021	-31166.8221	35636.2603	-0.87	0.3820
as.factor(Year)2022	2893.7621	35188.1775	0.08	0.9345
as.factor(Year)2023	60007.3440	35131.2196	1.71	0.0879*

and $\bar{R}^2 = 0.793$. Note the significance of the regressors for `2018` and `2023`.

We now incorporate state fixed effects:

$$\widehat{Claims\_Denied} = \hat{\beta}_0 + \hat{\beta}_1 \widehat{Claims\_Received} + (\text{time factors}) + (\text{state factors})$$

and arrive at a model with the following heteroskedasticity-robust regressors:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18182.4100	84040.4547	0.22	0.8287
Claims_Received	0.1540	0.0025	62.56	0.0000***
⋮	⋮	⋮	⋮	⋮
as.factor(Year)2018	-100022.0408	38274.2682	-2.61	0.0091*
⋮	⋮	⋮	⋮	⋮
as.factor(Year)2023	71485.6772	34269.4949	2.09	0.0372**
as.factor(State)AL	441105.6725	107413.8560	4.11	0.0000***
⋮	⋮	⋮	⋮	⋮
as.factor(State)FL	-169121.0821	88632.3272	-1.91	0.0566*
as.factor(State)GA	156490.3172	92908.9208	1.68	0.0924*
⋮	⋮	⋮	⋮	⋮
as.factor(State)NJ	366966.7190	112758.5459	3.25	0.0012***
⋮	⋮	⋮	⋮	⋮
as.factor(State)WY	-21354.4282	114316.4561	-0.19	0.8518

Adding state-fixed effects to our model produces a regression with an $\bar{R}^2 = 0.812$. This is a marked improvement over both models. Again, note the significance of the `2018` and `2023` regressors.

4 Model Development

4.1 Regression

We begin by attempting to estimate the number of denied claims by the various features of the dataset. To do so, we now impose a training and testing split of 75% and 25%, respectively. This allows us to circumvent potential over-fitting concerns.

4.1.1 Linear Regression

We begin with a simple ordinary least squares (OLS) regression model over $\log(Claims\_Denied)$ because inflexible models offer easy interpretability; and, as previously discussed, logarithmic transformations reduce heteroskedasticity in datasets. We thus fit a linear regression of

the following form:

$$\log(\widehat{Claims\_Denied}) = \hat{\beta}_0 + \hat{\beta}_1 \log(\widehat{Claims\_Received}) + (\text{time factors}) + (\text{state factors})$$

Which in turn gives the following heteroskedasticity-robust coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.8764700	0.1138376	7.6993	0.000***
log(*Claims_Denied*)	0.9896777	0.0123999	9.8131	0.000***
⋮	⋮	⋮	⋮	⋮
as.factor(Year)2018	-0.1492607	0.0426616	3.4987	0.000***
as.factor(Year)2019	-0.0688783	0.0402987	1.7092	0.088*
⋮	⋮	⋮	⋮	⋮
as.factor(Year)2021	-0.0759785	0.0386112	1.9678	0.049**
⋮	⋮	⋮	⋮	⋮
as.factor(Year)2023	0.0844709	0.0386352	2.1864	0.029**
⋮	⋮	⋮	⋮	⋮

with an $\bar{R}^2 = 0.901$ and $RMSE = 0.285$. We illustrate the fit of the OLS model in Figure 5.

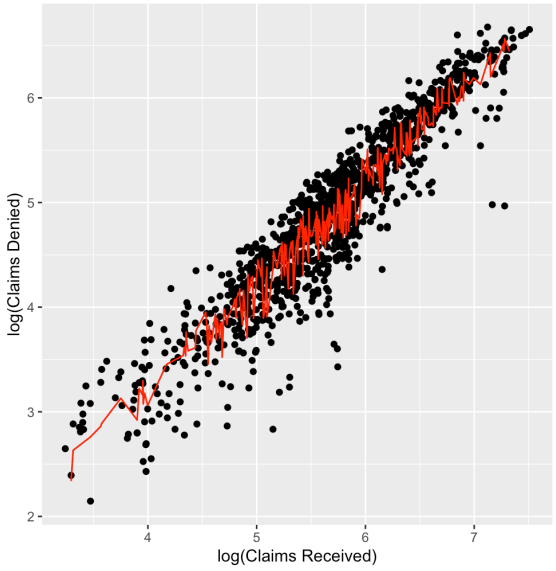


Figure 5: The result of OLS regression

This model implies significant time effects, but in both the upward and downward direction compared to the base case of 2015. That the direction changes challenges the assumption that insurance denials have been increasing overtime.

4.1.2 Support Vector Regression

Support Vector Regression (SVR) employs a Support Vector Machine (SVM) for regression tasks. As before, we fit the following model:

$$\begin{aligned}\log(\widehat{Claims\_Denied}) = \hat{\beta}_0 + \hat{\beta}_1 \log(\widehat{Claims\_Received}) \\ + (\text{time factors}) \\ + (\text{state factors})\end{aligned}$$

and now employ SVM tuning to determine an optimal $\gamma = 0.1$ and `cost` $= 1$.

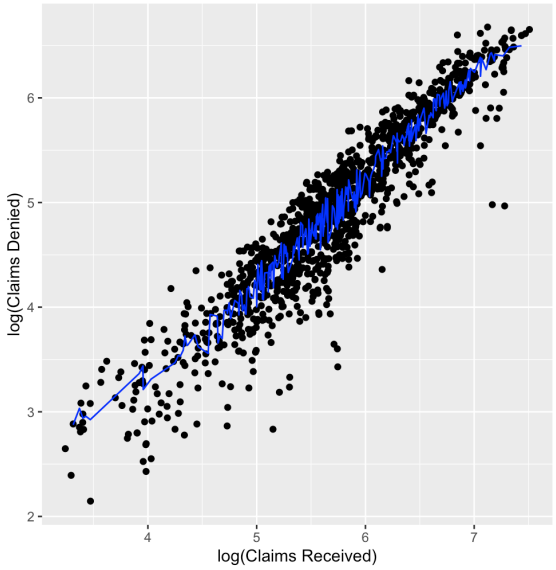


Figure 6: The result of an SVR fit

This method gives an $RMSE = 0.272$, an improvement over the linear model. Figure 6 indicates the graphical relationship between our SVR model and and our dataset. Despite the improvement in performance offered by the more flexible SVR, the loss of interpretability

compared to OLS makes this method of regression less valuable.

4.1.3 Random Forest

Random forests, an extension of the bagging method, are weighted collections of randomly generated decision trees generally used for classification. They may, nevertheless, be used in regression tasks and the added flexibility can offer marked improvements over OLS regressions.

We fit the same formula as in our SVR and OLS models, but arrive at a lower $RMSE = 0.367$. We plot the predicted values of $\log(\widehat{Claims\_Denied})$ against the true $\log(Claims\_Denied)$ in Figure 7.

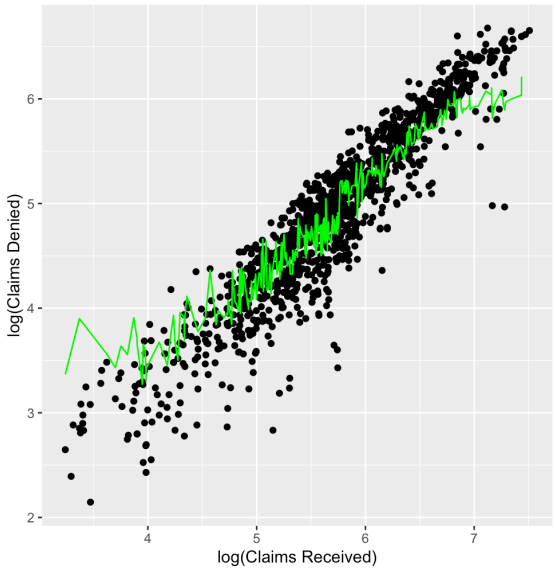


Figure 7: The result of a Random Forest regression

4.2 Ordinal Classification

We now assign the fourth quartile of `Denial_Rate` as `High_Denier` $= 1$, and all other data points as `High_Denier` $= 0$. This split is shown graphically in Figure 8.

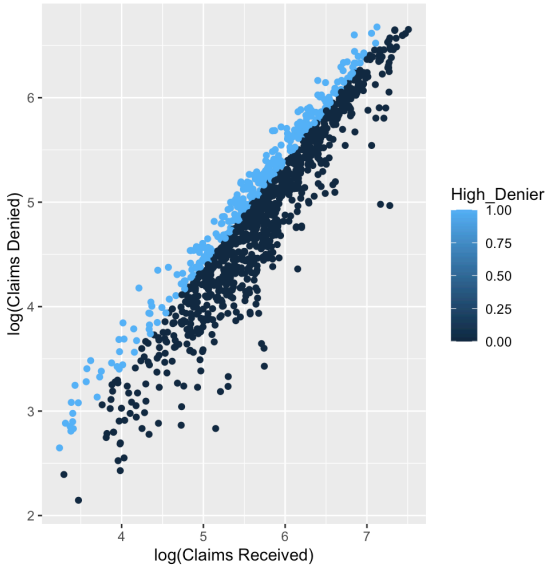


Figure 8: Ordinal splitting of the KFF datasets by denial rate

We attempt this classification task to verify if denial rates are independent of the number of claims when controlling for time and state effects.

4.2.1 Logistic Regression

Logistic regression models improve on the linear probability models by restricting the values the dependent variable can take to the range $P \in [0, 1]$. Despite this restriction, the relative inflexibility of logistic regression preserves interpretability.

We fit a model of the form

$$\begin{aligned} \widehat{P(High\_Denier = 1)} = \hat{\beta}_0 + \hat{\beta}_1 \log(\widehat{Claims\_Received}) \\ + (\text{time factors}) \\ + (\text{state factors}) \end{aligned}$$

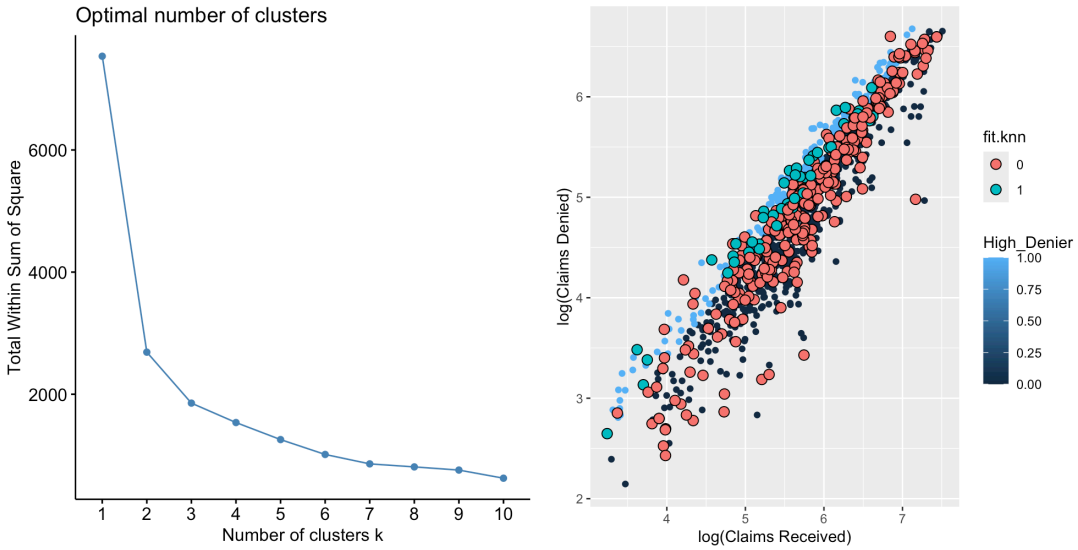
and arrive at a model with an AIC of 998.3 and a classification accuracy of 77.1%. As before, the coefficients on log(`Claims_Received`), `2018`, and `2023` are significant.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.29532	1.02554	1.263	0.206
logclaimed	-0.48040	0.11315	-4.246	0.000***
⋮	⋮	⋮	⋮	⋮
Year2018	-1.29940	0.47421	-2.740	0.006**
⋮	⋮	⋮	⋮	⋮
Year2023	0.85358	0.33806	2.525	0.012*
⋮	⋮	⋮	⋮	⋮

4.2.2 k-Nearest Neighbor

Classification by the k-nearest neighbor (kNN) method aims to find the dividing line between two classes. The smaller k is, the more flexible the model becomes, asymptotically approaching the Bayesian classifier. By convention, k is selected in the range of 5–10 or is determined optimally by the *elbow method*, which is done by visually inspecting a $WSS(k)$ graph. We restrict our model to log(`Claims_Denied`), log(`Claims_Received`), and `Year` and optimally determine $k = 11$ by the elbow method.

In so doing, we arrive at a classification accuracy of 81.6%.



5 Conclusions & Discussion

We employed a series of regression and classification methods to indicate the source of insurance claim denials and large denial rates. That we consistently see significant negative values for the regressor on 2018 and significant positive values for the regressor on 2023 suggests that the primary driver of denials is insurance uptake itself. With the repeal of the ACA individual mandate in 2018—and consequently the monetary penalty for not carrying health insurance—we see an overall decrease in the number of Americans carrying health insurance. In contrast, the availability of increased ACA subsidies starting in 2022 tracks with an increase in overall insurance uptake. That the signs of the regressor coefficients, a proxy for insurance uptake itself, track directionally with the number of denials suggests that, in aggregate, the number of insurance claims that are denied can be directly explained by the number of claims received. The strong $\bar{R}^2$ from our OLS regression further implies that health insurance providers by and large deny a fixed proportion of the claims they receive; moreover, at aggregate 90% of the variance in health insurance denials can be explained by the number of claims received when controlling for time and state fixed effects. The strength of this claim is demonstrated by the convergence of our regressions as illustrated in Figure 9.

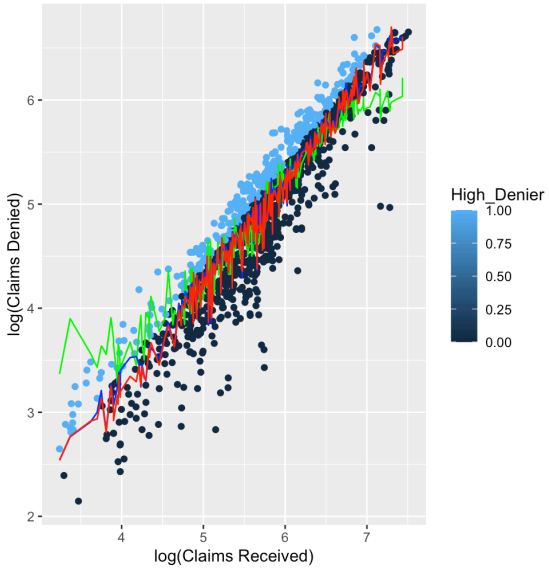


Figure 9: All regressions plotted

Although anecdotes of unfair insurance denials abound, we suggest that *the vast majority*

of insurance denials are fair. Despite time effects being significant, they track with insurance uptake overall. Research into insurance denials with greater granularity (e.g., having monthly denial information and having information on the procedures denied) is needed to demonstrate a systematic increase in denials over time not as a result of insurance uptake.

References

ARHQ. (2025, March). Information on the health status of americans, health insurance coverage, and access, use, and cost of health services. https://datatools.ahrq.gov/meps-hc/

Auerbach, D., Holtzblatt, J., Jacobs, P., Minicozzi, A., Moomau, P., & White, C. (2010). Will health insurance mandates increase coverage? synthesizing perspectives from health, tax, and behavioral economics. *National Tax Journal*, *63*(4.1), 659–679. https://doi.org/10.17310/ntj.2010.4.03

Baicker, K., Congdon, W. J., & Mullainathan, S. (2012). Health insurance coverage and take-up: Lessons from behavioral economics. *The Milbank Quarterly*, *90*(1), 107–134. https://doi.org/10.1111/j.1468-0009.2011.00656.x

CDC. (2024, June). U.s. uninsured rate drops by 26% since 2019. https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2024/20240618.htm

Dyer, O. (2025). Health insurers grapple with unitedhealthcare murder—and a public short on sympathy. *BMJ*, q2879. https://doi.org/10.1136/bmj.q2879

Justin Lo, M. L., & Wallace, R. (2025, January). Claims denials and appeals in aca marketplace plans in 2023. https://www.kff.org/private-insurance/issue-brief/claims-denials-and-appeals-in-aca-marketplace-plans-in-2023/

Karen Pollitz, J. L., & Wallace, R. (2023, February). Claims denials and appeals in aca marketplace plans in 2021. https://www.kff.org/private-insurance/issue-brief/claims-denials-and-appeals-in-aca-marketplace-plans/