Predicting Health Insurance Coverage of the U.S. Working Population Using Medical Expenditure Data in 2023

Abstract

Health insurance coverage is not universal in the United States. The working-age population experiences significant variation in coverage, owing to its dependence on income/employment status and membership in specific demographic groups. Given that insurance is a key determinant of healthcare access and well-being, we analyzed 2023 individual-level data from the Medical Expenditure Panel Survey Household Component (MEPS-HC) to identify key factors associated with health uninsurance. We employed a logistic regression model and found that Hispanic ethnicity, residence in the southern region, and marital status were significant predictors. Notably, having a usual place for medical care is associated with substantially lower odds of being uninsured, after controlling for income and other socioeconomic variables. Our findings inform policies that expand primary care access and outreach in the Southern region to close the persistent coverage gap among disadvantaged individuals.

I. Background and Significance

The United States operates under a complex healthcare system with multiple pathways to insurance, including publicly financed Medicaid and Medicare, as well as privately funded health insurance plans. However, coverage remains non-universal. Estimates for 2023 indicate that approximately 8% of the population remains uninsured (Bunch, 2023). In particular, a greater proportion of working-age adults (aged 18-64) were uninsured each year between 2020 and 2023. Unlike children and the elderly, these adults have coverage tied directly to employment and face restrictions on eligibility for publicly funded programs such as Medicare. In addition, research also suggests that systemic barriers, such as regional disparities and weak connections to the healthcare system, also play a key role. For example, nearly three-quarters (73.9%) of uninsured individuals reside in the South and West regions of the U.S., where Medicaid expansion has not yet been adopted (Tolbert, 2023). On the other hand, despite being eligible for public coverage, six in ten people who qualify for Medicaid or subsidized marketplace plans remain unenrolled (Drake, 2024), often due to reasons like information asymmetry, administrative complexity, and limited engagement with healthcare providers.

Motivated by these nuances, the research question we investigate is: “At the individual level, which variables predict uninsurance for the working age population (aged 18-64) in the U.S.?” We hope our model can help identify higher-risk groups and inform targeted interventions to improve access to care.

II. Data

The 2023 Medical Expenditure Panel Survey Household Component (MEPS-HC) includes 18,463 individuals from 8,133 families. It samples individuals initially selected through the National Health Interview Survey and provides information on insurance coverage, demographics, health conditions, and healthcare use, totaling 45 variables. The response variable is has no health insurance, a binary classification.

First, we limited our sample to the working-age population (aged 18-64 years). We excluded observations with missing demographic information (legal marital status and race) because imputing these data would be unethical. For other variables, we performed logistic- or multinomial regression-based imputation using variables with complete data (e.g., age, sex, region, marital status, race, income, etc.). We also identified two variables with high proportions of “Not Applicable.” This indicated that the question was not asked of the individual due to the survey design. For example, the variable student status had 88% “Not Applicable” because the question was asked only of people aged 17-23. Imputing would be inaccurate, as the current dataset lacks student status for individuals aged 23+, and assumptions, such as treating all individuals with the highest education of undergraduate or graduate as no longer students, risk missing actual graduate and non-traditional students. Similarly, being advised to quit smoking in the past 12 months had 93% “Not Applicable”, since it is only asked of smokers who have visited a doctor in the past 12 months. On the other hand, the variable smoke now provides better coverage because the question is asked of individuals aged 18 years and older, which aligns with our population of interest. As a result, we dropped both variables.

After data cleaning, 28 predictors remained in the dataset. Given that the current dataset violates the independence assumption required for logistic regression, as it includes multiple individuals from the same household, we implemented random sampling within households,

reducing the sample size to 6,195. Finally, we assessed multicollinearity using VIF with a threshold of 10, with all predictors' VIF values well below 3. Hence, we retained all predictors.

III. Methods and Results

We employed backward stepwise regression using AIC and BIC to select the best first-order model, given its computational efficiency. The selection procedure yielded two candidate models; we evaluated their predictive performance using cross-validation to determine the best first-order model. The primary performance metric was the area under the ROC curve (AUC), which assesses model performance across all possible discrimination thresholds. In 5-fold cross-validation, both models yielded a mean CV score of approximately 0.85, as indicated by similar area under the ROC curves. Given the sparse (593 uninsured) dataset, we tuned the discrimination threshold to identify the value that best balances sensitivity (true-positive rate) and specificity (true-negative rate). Through 5-fold cross-validation, we evaluated sensitivity and specificity across thresholds ranging from 0.1 to 0.5. At the threshold of 0.1, both models yielded high sensitivity (0.80) at the cost of reduced, but still moderate, specificity (0.76). Because our primary goal is to identify uninsured individuals, prioritizing sensitivity is appropriate in this context.

Since the two models had similar predictive performance, we selected the BIC model as our best first-order model, given that it is more parsimonious, yields more precise estimates of regression coefficients, and permits greater degrees of freedom in statistical inference. We then explored the possibility of interaction terms using the same model-selection method. The cross-validated AUC scores of the candidate interaction models were 0.86 (BIC) and 0.87 (AIC). As a result, we kept the best first-order model as our final model, for the same reason of parsimony.

After finalizing the model, we proceeded with model diagnostics. We focused on outlying observations in the dataset and investigated whether their inclusion would violate key assumptions of regression. The four residual diagnostic plots highlight six outlying observations (residual case numbers 16614, 15254, 8397, 6876, 5450, and 774). We found that all six individuals reported having no health insurance and high healthcare utilization (e.g., annual total number of visits made to office-based medical providers and count of all prescribed medications purchased in 2023). However, these individuals' health conditions may help explain the unusual pattern. For example, the individual who reported the highest number of prescribed medications (case 16614) is diabetic, reflecting medical needs for a chronic condition; the people who reported high numbers of office-based medical visits (cases 774 and 6876) both report having had cancer before. Lastly, the person who reported 100 office-based medical visits (case 5450) may have a chronic health condition not captured in the data. In general, we could not confidently conclude that these patterns are attributable to a coding or a data-collection issue.

To address that, we calculated delta deviances to determine whether they are truly influential. We plotted all delta deviances against the residual cases to identify spikes, indicating significant changes in residual deviance when an observation is removed. The distribution does not show a noticeable spike despite the six cases having higher delta deviances. Thus, we set a threshold for the delta deviance at 11, exclude these values, and refit the model to assess whether predictive performance changes significantly. The resulting CV score was 0.86, which is essentially the same as that of the entire dataset, thereby motivating us to retain it for analysis.

In our final model, levels “Northeast” and “West” in the region variable and “Some college” in the highest level of education completed variable are conditionally insignificant. The p-values for the likelihood ratio tests for both predictors were near 0, indicating that we should not remove them. Therefore, we combined “Northeast”, “West”, and “Midwest” and created a dummy variable indicating whether the region is in the South. Similarly, we merged “Some college” with “College graduate” (the baseline) to generate the baseline level “Any college”.

Our final equation yielded: logit($\hat{P}$(Uninsured)) = -1.638 + 0.01948 (Age) + 0.9068 (South) - 0.4023 (Married (Spouse in Household)) - 0.9629 (Not Hispanic) - 0.6654 (Graduate Education) + 0.5082 (High School Graduate) + 0.9407 (Less Than High School) - 0.000007658 (Personal Income) - 0.5932 (SNAP Receipts) - 1.130 (Has Usual Place for Healthcare) - 0.04576 (Annual Total Number of Visits to Office-Based Medical Providers) - 0.05953 (Annual Count of Prescribed Medications Purchased). All predictor variables in our final equation were statistically significant at the 0.05 significance level. The likelihood ratio test for goodness of fit yielded a p-value of 1, suggesting no sufficient evidence of lack of fit relative to the saturated model.

IV. Discussion and Future Considerations

We exponentiated the estimates and interpreted them as multipliers of the predicted odds. We found that has usual place for healthcare had the largest magnitude, with individuals who have a usual place for medical care having about 0.32 times the odds of being uninsured compared to individuals without a usual place for care, after adjusting for simultaneous linear changes in the other predictors. This indicates the importance of connection to the healthcare system, even after controlling for socioeconomic status. Our model also confirmed that the predicted odds of being uninsured are higher for individuals in the South region. As of 2025, 6 of the 10 states that have not adopted Medicaid expansion are in the South region defined by MEPS-HC (Alabama, Florida, Georgia, South Carolina, Tennessee, and Texas). Moreover, the predicted odds of being uninsured are higher for Hispanic individuals, possibly due to their overrepresentation in sectors with low rates of employer-sponsored health insurance. Lastly, the predicted odds of being uninsured are lower among married individuals with a spouse in the household, suggesting that geographic separation may affect married individuals without a spouse in the household.

There are some limitations in our study. Our dataset doesn’t include current employment status, one of the strongest predictors of insurance. The ever worked in 2023 variable is a binary classification and does not provide job characteristics, making it an inadequate proxy. Additionally, MEPS-HC omits health condition questions for individuals in certain survey rounds. We assumed that timing should not be correlated with demographics or health status and imputed it as a missing value. However, if this were not the case, imputation may introduce bias. Lastly, MEPS-HC consistently over-samples certain population groups, especially minorities. As a result, the sample may not be fully representative of the U.S. working-age population.

To improve predictions of health insurance status, we may examine corrections to the survey design using MEPS-HC survey weights. To validate our imputation assumption, we obtain more detailed information on survey timing and skip patterns. Lastly, we may consider undersampling approaches during model selection to address data sparsity.

Appendix

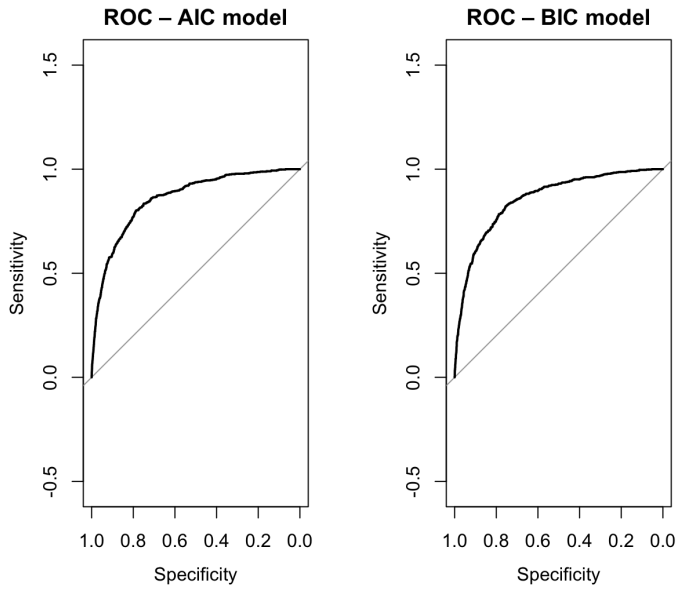
- **Variable chart**

Variable	Type	Levels
AGE	Numeric	18–64
ANYLMT	Binary	Any limitation
ASTHMAEV	Binary	Ever asthma
CANCEREV	Binary	Ever cancer
CHEARTDIEV	Binary	Coronary heart disease
CHOLHIGHEV	Binary	High cholesterol
DIABETICEV	Binary	Diabetes
EDUCYR	Categorical	<HS / HS / Some college / College / Grad
ERTOTVIS	Count	ER visits
EXERMODVIG5D	Binary	Exercise 5x/week
FOODSTYN	Binary	SNAP Yes/No
FTOTVAL	Numeric	Family income
HEARTATTEV	Binary	Heart attack
HEARTCONEV	Binary	Heart condition
HINOTCOV	Binary	No insurance
HISPETH	Categorical	Hispanic / Not Hispanic
INCTOT	Numeric	Income total
MARSPOUIN	Categorical	Spouse present (Y/N/NA)
MARSTAT	Categorical	8 marital statuses
OBTOTVIS	Count	Office visits
RACEA	Categorical	White / Black / Asian/PI / AIAN / Multiple
REGIONMEPS	Categorical	NE / MW / S / W
RXPRMEDSNO	Count	Medications
SEX	Categorical	Male / Female
SMOKENOW	Binary	Current smoker
SOUTH	Binary	South / Other
STROKEV	Binary	Stroke
USUALPL	Binary	Has usual place
WORKEV	Binary	Ever worked

- **VIF scores of predictors**

```
> car::vif(full_model)
                  GVIF Df GVIF^(1/(2*Df))
AGE           1.653717  1        1.285969
SEX           1.165183  1        1.079436
MARSTAT      10.886869  4        1.347761
REGIONMEPS    1.196846  3        1.030401
MARSPOUIN     7.757290  1        2.785191
RACEA         1.440097  4        1.046644
HISPETH       1.440020  1        1.200008
EDUCYR        1.367468  4        1.039895
WORKEV        1.185622  1        1.088863
INCTOT        1.839560  1        1.356304
FTOTVAL       1.664699  1        1.290232
FOODSTYN      1.100060  1        1.048838
USUALPL       1.133722  1        1.064764
ANYLMT        1.193373  1        1.092416
ASTHMAEV      1.038210  1        1.018926
CANCEREV      1.049301  1        1.024354
CHEARTDIEV    1.128186  1        1.062161
CHOLHIGHEV    1.239928  1        1.113521
DIABETICEV    1.216230  1        1.102828
HEARTATTEV    1.086251  1        1.042233
HEARTCONEV    1.054203  1        1.026744
STROKEV       1.104180  1        1.050800
SMOKENOW      1.098127  1        1.047915
EXERMODVIG5D  1.054270  1        1.026776
OBTOTVIS      1.212616  1        1.101189
ERTOTVIS      1.059720  1        1.029427
RXPRMEDSNO    1.442765  1        1.201152
```

- **ROC curves**



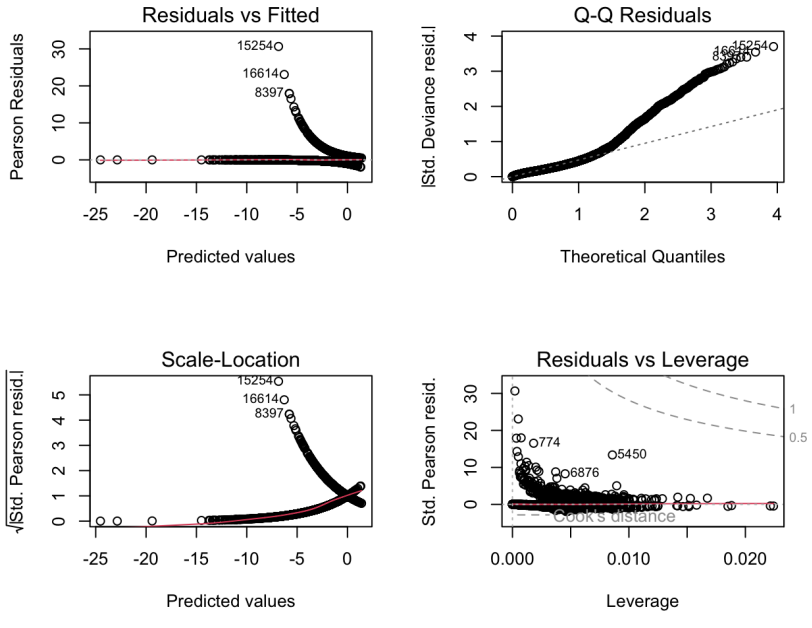
- **Cross-validated sensitivity and specificity under different discrimination thresholds**

```
> data.frame(thresholds, CV_sensitivity_AIC, CV_specificity_AIC)
  thresholds CV_sensitivity_AIC CV_specificity_AIC
1       0.10          0.8105211          0.7661318
2       0.15          0.6869288          0.8457255
3       0.20          0.6067912          0.8898119
4       0.25          0.5348039          0.9237424
5       0.30          0.4437253          0.9469499
6       0.35          0.3666742          0.9631948
7       0.40          0.3029887          0.9742682
8       0.45          0.2276552          0.9817561
9       0.50          0.1707118          0.9874853
> data.frame(thresholds, CV_sensitivity_BIC, CV_specificity_BIC)
  thresholds CV_sensitivity_BIC CV_specificity_BIC
1       0.10          0.8047894          0.7648722
2       0.15          0.6914805          0.8425115
3       0.20          0.6220303          0.8903522
4       0.25          0.5316543          0.9223084
5       0.30          0.4374966          0.9451665
6       0.35          0.3475769          0.9621210
7       0.40          0.2741020          0.9731897
8       0.45          0.2125452          0.9828400
9       0.50          0.1691280          0.9887269
```

- **CV score table**

```
        Model    CV_AUC
1     lm_AIC 0.8526433
2     lm_BIC 0.8522329
3 lm_INT_AIC 0.8655127
4 lm_INT_BIC 0.8554135
```

- **Regression diagnostic plots**

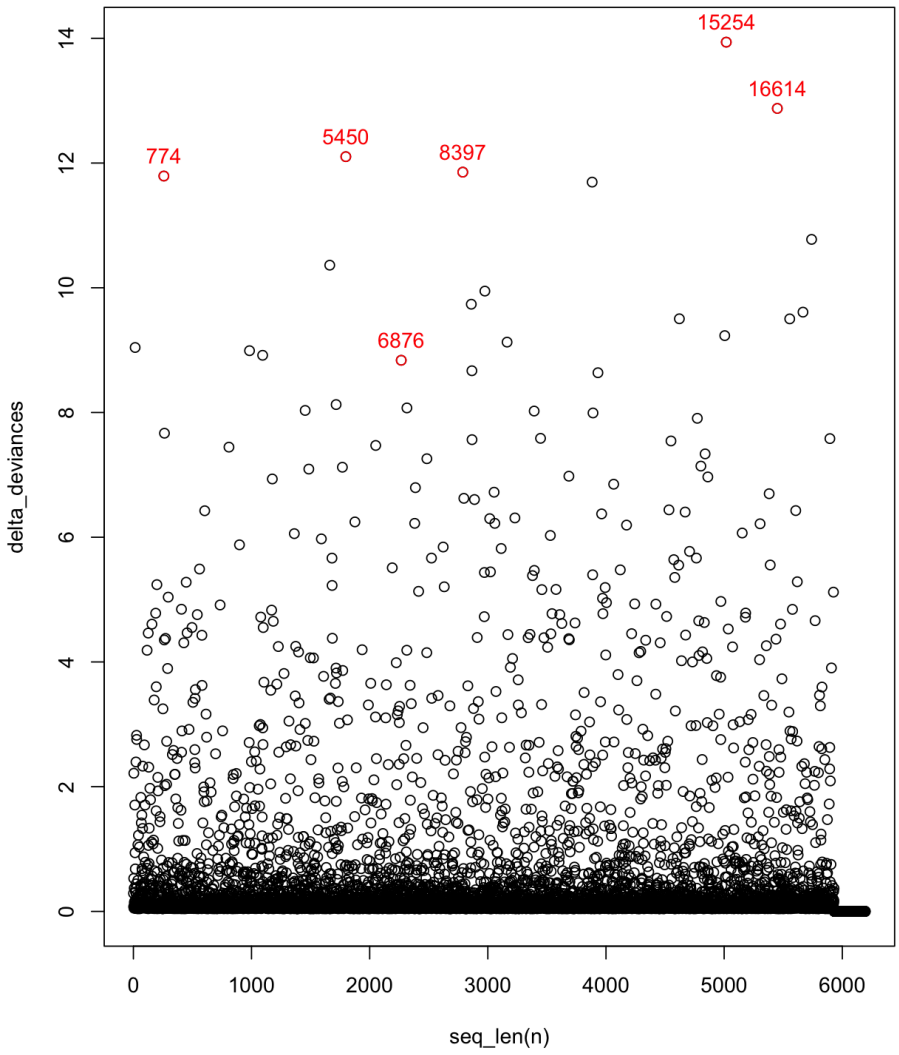


- **Data rows for outlying observations**

	AGE	SEX	MARSTAT	REGIONMEPS	MARSPOUIN	RACEA	HISPETH	EDUCYR	WORKEV	INCTOT	FTOTVAL	FOODSTYN	USUALPL	HINOTCOV	ANYLMT
8397	48	Female	Married	South	Spouse in household	White	Not Hispanic	Graduate education	Yes	59500	124484	No	Yes	Yes	No
15254	41	Female	Married	Midwest	Spouse in household	White	Not Hispanic	Graduate education	Yes	80000	119000	No	Yes	Yes	No
16614	61	Female	Married	South	Spouse in household	Black	Not Hispanic	College graduate	Yes	0	38461	Yes	Yes	Yes	Yes
774	62	Female	Married	West	Spouse in household	White	Not Hispanic	High school graduate	Yes	28680	28680	No	Yes	Yes	No
5450	61	Female	Separated	South	Not married / no spouse in household	White	Not Hispanic	Some college	Yes	37396	37396	No	No	Yes	No
6876	63	Female	Never married	South	Not married / no spouse in household	White	Not Hispanic	College graduate	Yes	57133	93373	No	No	Yes	Yes

	ASTHMAEV	CANCEREV	CHEARTDIEV	CHOLHIGHEV	DIABETICEV	HEARTATTEV	HEARTCONEV	STROKEV	SMOKENOW	EXERMODVIGSD	OBTOTVIS	ERTOTVIS	RXPRMEDSNO
8397	Yes	No	No	No	No	No	No	No	No	No	39	0	10
15254	No	No	No	No	No	No	No	No	No	Yes	33	0	13
16614	No	No	No	Yes	Yes	No	No	Yes	Yes	No	4	0	55
774	No	Yes	No	Yes	No	No	No	No	No	No	58	0	5
5450	No	No	No	No	No	No	No	No	Yes	Yes	100	0	0
6876	No	Yes	No	No	No	No	No	No	No	No	48	0	17

- **Delta deviance plot**



- **CV score table for the full and trimmed datasets**

	Model	CV_AUC
1	lm_BIC	0.8522329
2	lm_BIC_TRIMMED	0.8615222

- **Summary output for final model**

```
> final_model_combined <- glm(HINOTCOV ~ AGE + SOUTH + MARSPOUIN + HISPETH + EDUCYR + INCTOT + FOODSTYN + USUALPL + OBTOTVIS + RXPRMEDSNO, family = binomial, data = meps_sampled)
> summary(final_model_combined)

Call:
glm(formula = HINOTCOV ~ AGE + SOUTH + MARSPOUIN + HISPETH + 
    EDUCYR + INCTOT + FOODSTYN + USUALPL + OBTOTVIS + RXPRMEDSNO, 
    family = binomial, data = meps_sampled)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.638e+00  1.969e-01  -8.322  < 2e-16 ***
AGE                          1.948e-02  3.875e-03   5.026 5.01e-07 ***
SOUTHSouth                   9.068e-01  9.879e-02   9.179  < 2e-16 ***
MARSPOUINSpouse in household -4.023e-01  1.107e-01  -3.635 0.000278 ***
HISPETHNot Hispanic          -9.629e-01  1.015e-01  -9.491  < 2e-16 ***
EDUCYRGraduate education     -6.654e-01  2.646e-01  -2.515 0.011913 *  
EDUCYRHigh school graduate    5.082e-01  1.192e-01   4.263 2.02e-05 ***
EDUCYRLess than high school   9.407e-01  1.410e-01   6.673 2.51e-11 ***
INCTOT                       -7.658e-06  1.572e-06  -4.872 1.11e-06 ***
FOODSTYNYes                  -5.932e-01  1.518e-01  -3.907 9.35e-05 ***
USUALPLYes                   -1.130e+00  1.054e-01 -10.719  < 2e-16 ***
OBTOTVIS                     -4.576e-02  1.252e-02  -3.655 0.000257 ***
RXPRMEDSNO                   -5.953e-02  9.406e-03  -6.329 2.47e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3910.0  on 6194  degrees of freedom
Residual deviance: 2887.3  on 6182  degrees of freedom
AIC: 2913.3

Number of Fisher Scoring iterations: 7
```

Works Cited

Bunch, L. (2024, September 10). *While the share of uninsured remained at about 8% in 2023, rates varied by age and Poverty Level*. Census.gov. https://www.census.gov/library/stories/2024/09/health-insurance-coverage.html

Tolbert, J., Bell, C., Cervantes, S. & Damico, A. (2025, August 12). *How many uninsured are in the coverage gap and how many could be eligible if all states adopted the Medicaid expansion?*. KFF. https://www.kff.org/medicaid/how-many-uninsured-are-in-the-coverage-gap-and-how-many-could-be-eligible-if-all-states-adopted-the-medicaid-expansion/

Drake, P. (2025, August 9). *A closer look at the remaining uninsured population eligible for Medicaid and CHIP*. KFF. https://www.kff.org/uninsured/a-closer-look-at-the-remaining-uninsured-population-eligible-for-medicaid-and-chip/?utm_source=chatgpt.com