PRENATAL CARE & ITS IMPACT ON HEALTHY BIRTHS

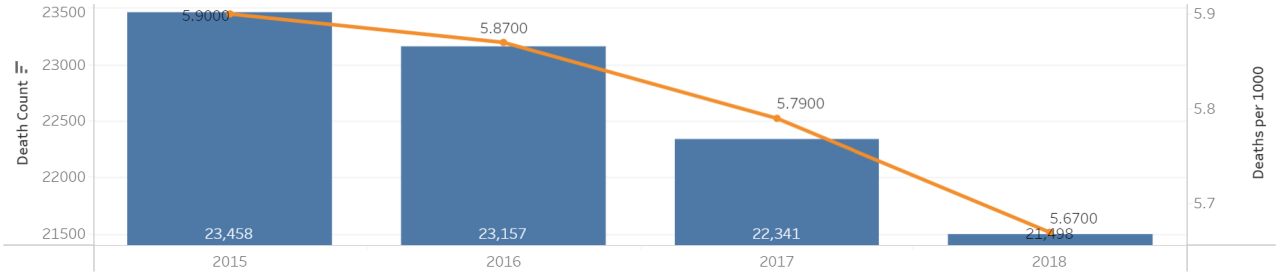
OCTOBER 2021

DS4A Women's - Team 20

Ariane Erickson, Daniella Furman, Yerin Lim, Kelsey Maass, Vi Nguyen, Jialing Wang, Sophia Yang

TABLE OF CONTENTS

INTRODUCTION



01 Trends in infant death counts
U.S. infant deaths per 1,000 live births, 2015-2018.

> *"...with respect to infant mortality, the U.S. ranks 33 out of 36 Organization of Economic Cooperation and Development (OECD) nations. In 2018, while infant mortality reached an all-time low in the U.S. [...] still more than 21,000 infants died. Compared to countries with a similar GDP, the U.S. infant mortality rate is much higher. France and the U.K., for example, have 3.8 deaths per 1,000 live births." (Cohen, 2021)*

Although the United States is a wealthy and developed country, it has higher rates of infant morbidity and mortality when compared to countries with a similar GDP. Within the U.S., neonatal, post-neonatal, and infant morbidity and mortality rates vary considerably by socioeconomic status, geographic location, and maternal race. For example, in 2013, mortality rates across these categories were approximately two times higher for children of non-Hispanic black women than for children of non-Hispanic white women (Singh, 2021), while white, college-educated women had mortality rates comparable to similar demographics in European nations (Cohen, 2021).

In this project, we set out to understand the various factors that contribute to infant morbidity and mortality. After examining various data sets related to both maternal and infant outcomes, we selected the CDC Natality and Linked Birth / Infant Death datasets to analyze. Our initial exploratory data analysis (EDA) revealed many benefits of increasing numbers of prenatal visits, which aligned with the understanding that access to and use of prenatal care can reduce pregnancy complications due to pre-existing conditions, reduce the risk of fetal/infant complications, and identify aberrations in the normative development of the fetus. For the majority of births in the U.S. (77%), prenatal care starts during the first trimester (Martin et al., 2019); however, disparities in access to early prenatal care may contribute to variations in birth outcomes.

Using a combination of modeling techniques, the team explored the following questions:

- How does prenatal care affect birth outcomes?
- What factors influence who receives prenatal care?

Overall, we observed that those with access to prenatal care had lower infant death rates, lower rates of admission to the neonatal intensive care unit (NICU), and a lower percentage of infants with very low birth weight. Furthermore, starting prenatal care earlier in the pregnancy was a significant factor in lowering death rates, with a greater impact for infants born at earlier gestational age. Finally, we found that differences in insurance status, race, and education were important indicators in who receives prenatal care.

DATA ANALYSIS & COMPUTATION

DATA SETS + DATA WRANGLING & CLEANING

Data Sets

Our primary data sources were the CDC Natality and CDC Linked Birth / Infant Death Records. These data sets were accessed from the CDC Wonder query system for aggregated results and from the NBER Public Use Data Archive for non-aggregated results.

Data Set	Source
CDC Natality Records The Natality online databases report counts of live births occurring within the United States to U.S. residents. Counts can be obtained by a variety of demographic characteristics, such as state and county of residence, mother's race, and mother's age, and health and medical items, such as tobacco use, method of delivery, and congenital anomalies. The data are derived from birth certificates.	• CDC Wonder • NBER Vital Statistics
CDC Linked Birth / Infant Death Records This data collection provides counts and rates for deaths of children under 1 year of age, occurring within the United States to U.S. residents. Information from death certificates has been linked to corresponding birth certificates. Data are available by county of mother's residence, child's age, underlying cause of death, gender, birth weight, birth plurality, birth order, gestational age at birth, period of prenatal care, maternal race and ethnicity, maternal age, maternal education and marital status. The data are produced by the National Center for Health Statistics.	• CDC Wonder

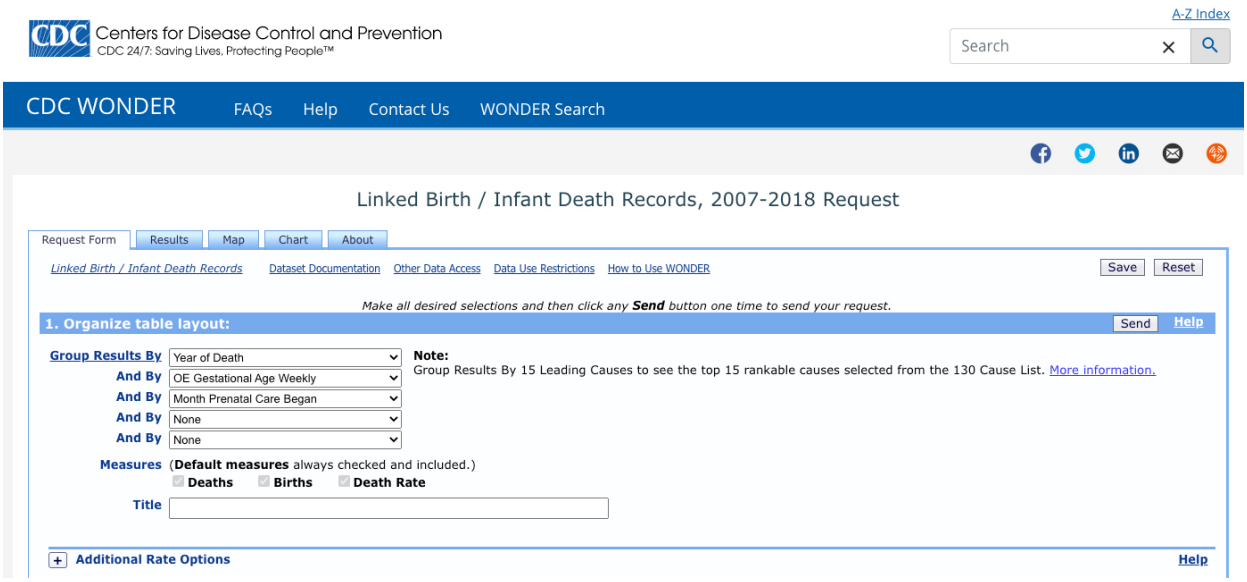
Our initial EDA looked into relationships among variables in both the aggregated and non-aggregated data sets. When using aggregated data, we additionally had access to geographical information. For the majority of our statistical data analysis (SDA), we focused on the non-aggregated natality records, which did not include geographic information to protect privacy.

DATA ANALYSIS & COMPUTATION

DATA SETS + DATA WRANGLING & CLEANING

Data Wrangling & Cleaning

Aggregated data from CDC Wonder was accessed through their online query system. Here users can request data grouped by different variables and filtered by specific values, and results can be saved as text files. The size of each data set we used varied depending upon the categories and ranges queried.



02 CDC Wonder
Example of a CDC Wonder query for birth / infant death data grouped by year of death, gestational age, and month prenatal care began.

Non-aggregated natality data was downloaded from the NBER Public Use Data Archive as CSV files. These data sets contained individual records for all births that occurred within a given year. We focused on the 2018 data set, which contained 3,801,534 entries and 240 variables. To decrease the number of records, we chose to consider only single births (i.e., no twins or triplets) that corresponded to a mother's first pregnancy. These initial filters reduced our data set to 1,146,108 records.

DATA ANALYSIS & COMPUTATION

DATA SETS + DATA WRANGLING & CLEANING

Through our EDA, we identified complications inherent in addressing both gestational age at birth and month prenatal care began, so we decided to only include births that occurred at or after 28 weeks (roughly the third trimester). Additionally, all of our data sets contained missing values or entries flagged as unreliable, which we excluded from our analysis. For example, in our analysis of who receives prenatal care, we excluded births that occurred after 40 weeks due to data sparsity.

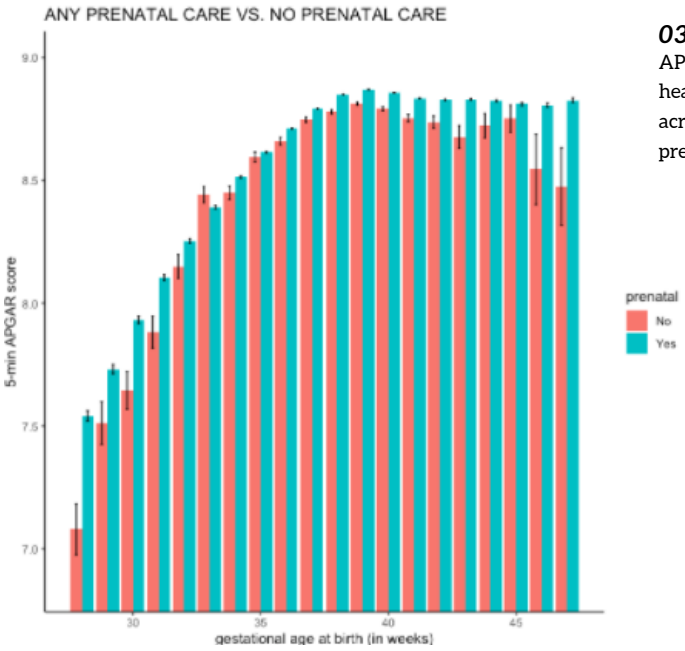
Trimester	Month	Week
1st	1	1 - 4
	2	5 - 8
	3	9 - 13
2nd	4	14 - 17
	5	18 - 22
	6	23 - 27
3rd	7	28 - 31
	8	32 - 35
	9	36 - 40

We also performed various aggregations, transformations, and filters to define new variables. For example, in our EDA we used birth, death, and population counts to calculate birth and death rates. We also computed the trimester that prenatal care began by grouping entries by month or week prenatal care began (see table). In our SDA, we transformed categorical variables into 0-1 indicators, we centered model inputs, we explored record-matching to control for the influence of specific variables, and we created balanced subsets of our data to better predict rare outcomes (explained in more detail in our SDA section).

DATA ANALYSIS & COMPUTATION

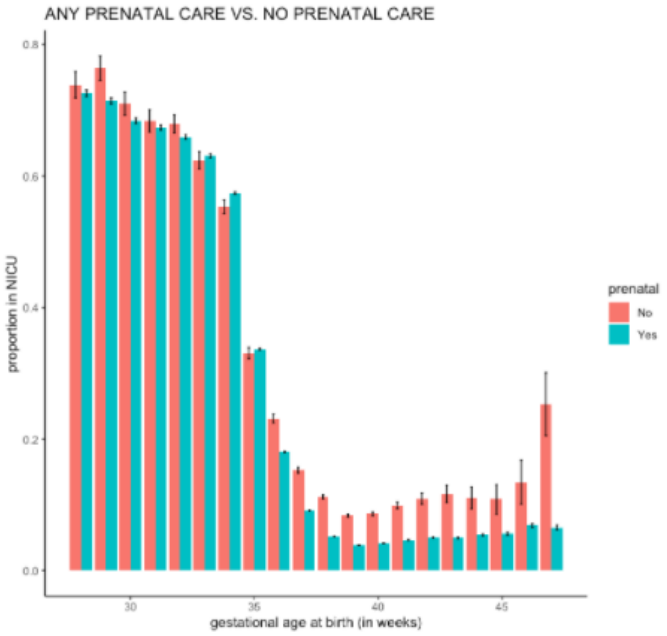
EXPLORATORY DATA ANALYSIS

To familiarize ourselves with the CDC data sets, we looked for trends among the included variables. The following visualizations are a snapshot of our exploration into prenatal care and birth outcomes. We also looked into geographic trends, with examples in our prenatal care and infant death dashboards.



03 APGAR score

APGAR score (a quick metric used to evaluate the health of newborns 5 minutes after birth) is higher across nearly all gestational ages (28+ weeks) when prenatal care is provided. [2018 data]



04 NICU admission

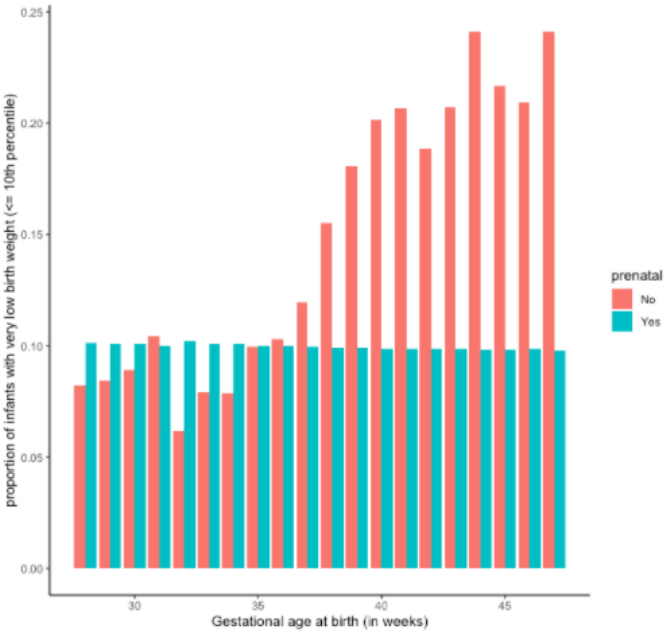
Prenatal care additionally reduces the proportion of full-term (37+ weeks) infants transferred to the newborn intensive care unit (NICU). [2018 data]

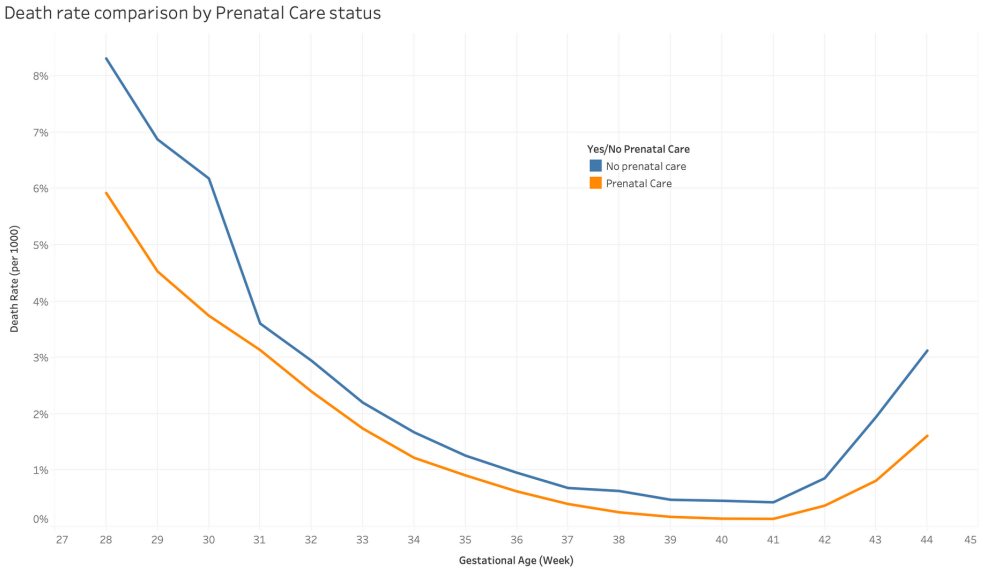
DATA ANALYSIS & COMPUTATION

EXPLORATORY DATA ANALYSIS

05 Very low birth weight

Mothers who don't receive prenatal care are more likely to give birth to infants who are small for gestational age after 37 weeks. Prenatal monitoring is likely critical for identifying babies who are at risk for being small for gestational age or who have other complications. [2018 data]



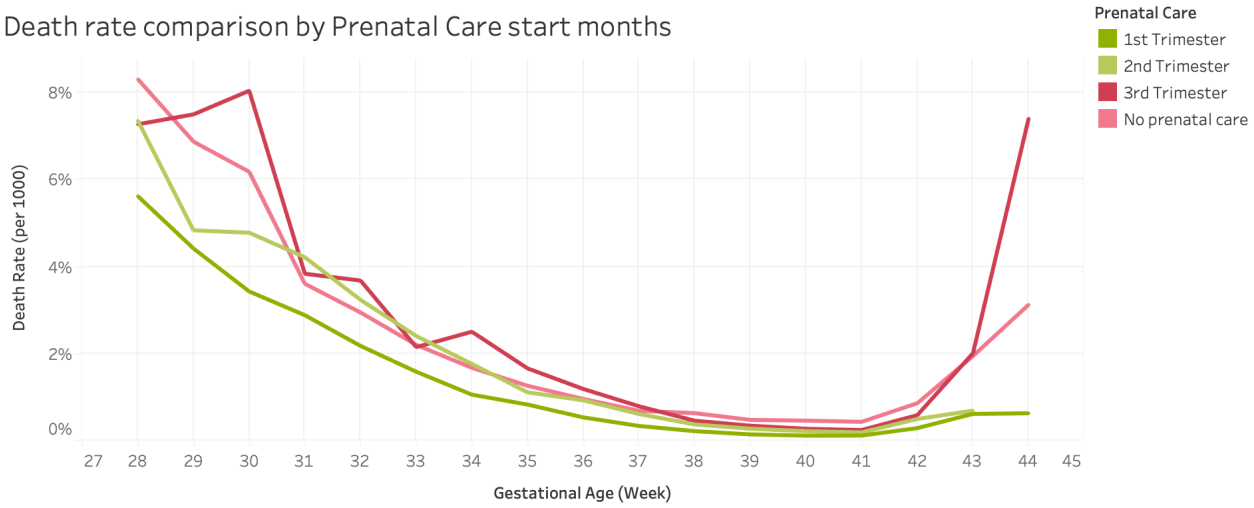


06 Death rate and prenatal care

The gap in death rate between infants whose mothers did or did not receive prenatal care is the highest when born at earlier gestation; this suggests that not receiving prenatal care has a large impact on infant survival for babies born earlier.

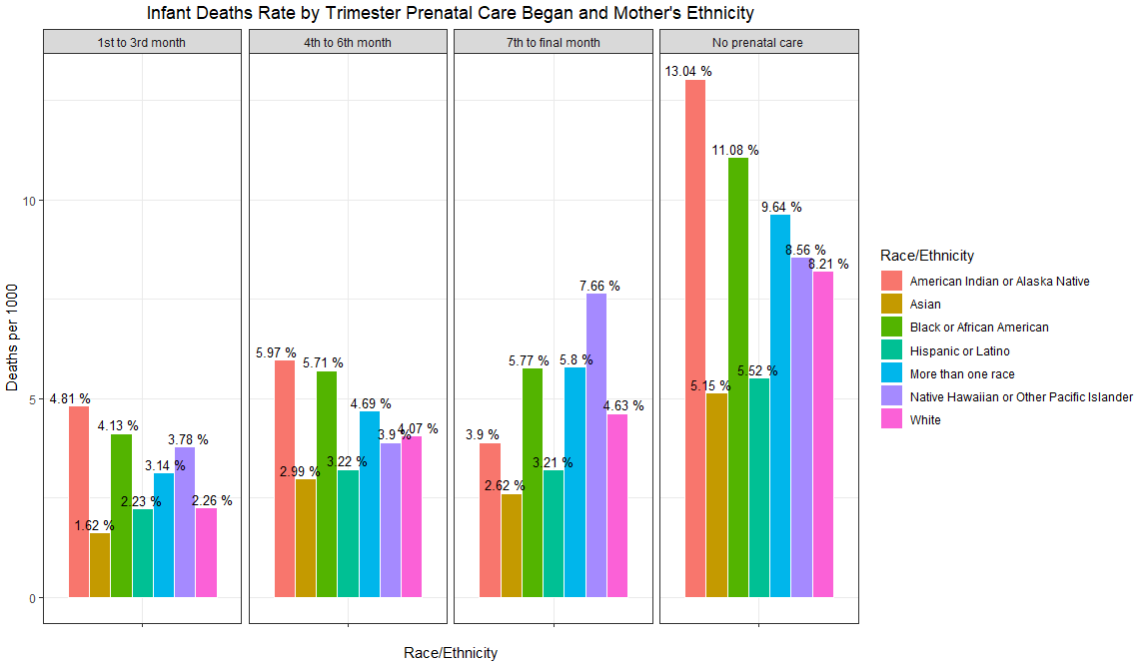
DATA ANALYSIS & COMPUTATION

EXPLORATORY DATA ANALYSIS



07 Death rate and trimester care began

Across gestational age, we see a difference in infant death rates as a function of when prenatal care began. The gap in death rates between mothers who started prenatal care in the 1st or 2nd trimester and those who did not receive prenatal care is ~2.7%.

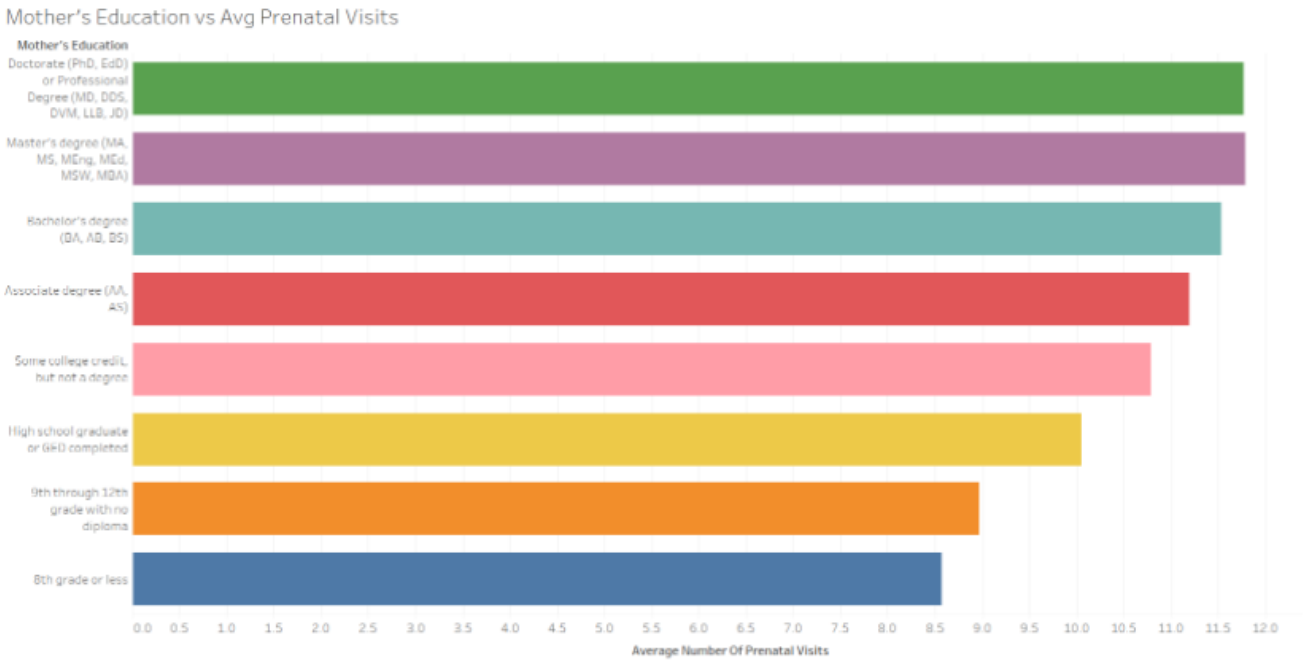


08 Mother's race/ethnicity

While the infants of White, Hispanic, and Asian mothers have the lowest overall rate of death, mothers starting prenatal care in the first trimester (1st-3rd month) experienced lower rates of infant death overall.

DATA ANALYSIS & COMPUTATION

EXPLORATORY DATA ANALYSIS



09 Mother's education

Mothers with higher levels of educational attainment have a greater number of prenatal visits, with an average of 12 prenatal visits for mothers with a master's or doctorate-level degree vs. only 8.5 average visits for mothers who received an 8th grade education or lower.

DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

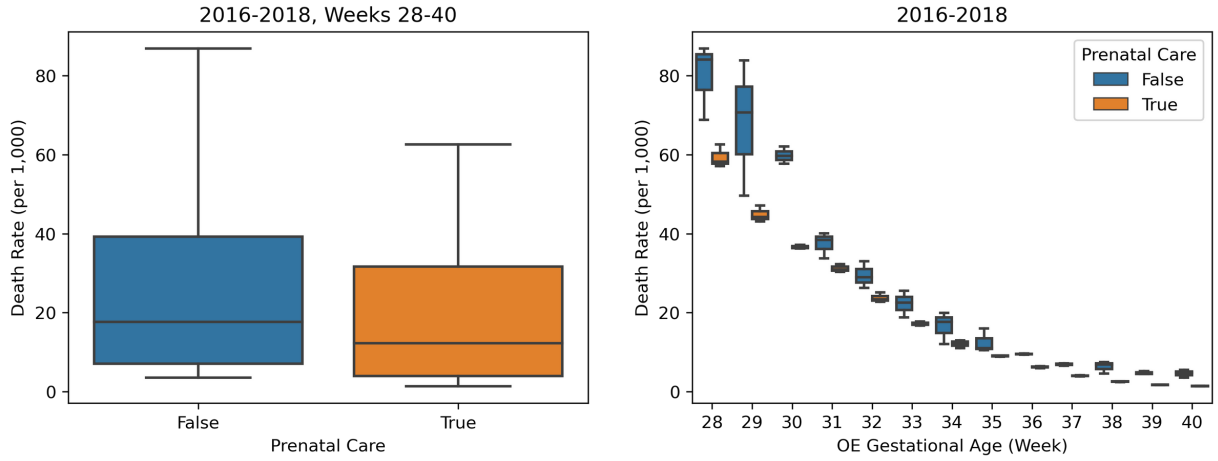
Two key observations from our EDA were that prenatal care could lead to improved birth outcomes, and that the percentage of women who received care varied among different demographics. To further investigate these patterns, we focused our SDA on the following two questions:

1. How does prenatal care affect birth outcomes?
2. What factors influence who receives prenatal care?

How does prenatal care affect birth outcomes?
PART I: Death rate

The first birth outcome we considered was infant death rate, defined in the Linked Birth / Infant Death Records dataset as the number of deaths of children under 1 year of age per 1,000 live births. Due to missing values and changes in reporting methods over time, we focused on a subset of the 2007-2018 data set including data from 2016-2018.

Below, we see how death rate varies based on whether or not the mother received prenatal care. For all infants born at 28-40 weeks gestation, the average death rate for mothers who did not receive prenatal care was 27.55 deaths per 1,000 live births, higher than the average rate of 19.23 for the mothers that did receive care (left). Additionally, this rate varies by gestational age. We observed that not only was the rate higher without prenatal care for each individual weeks gestation, but that the difference was greater for infants born at earlier gestation (right). To test if this relationship was statistically significant, we fit a log-linear model to the death rate based on prenatal care and gestational age.



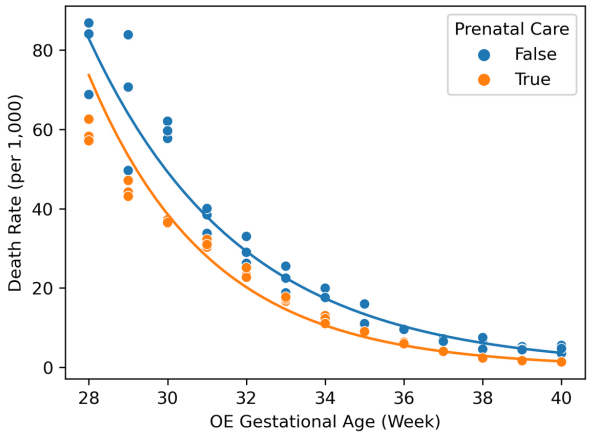
10 Infant death rates

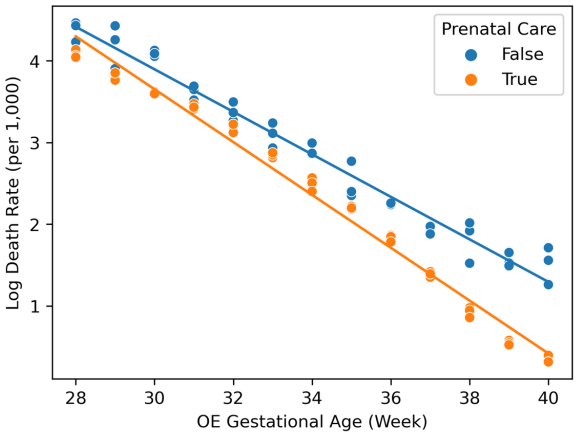
Infant death rates for 2016-2018 based on prenatal care and gestational age in weeks. Receiving prenatal care is associated with lower death rates, with larger differences observed for infants born at earlier gestation.

DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

All of the coefficients in our model were statistically significant. As expected, both utilization of prenatal care and a higher gestational age were associated with lower death rates. Furthermore, the reduction in death rate with advancing gestational age was greater for mothers who received prenatal care.





Dep. Variable:	Log Death Rate	**R-squared:**	0.980
Model:	OLS	**Adj. R-squared:**	0.979
Method:	Least Squares	**F-statistic:**	1218.
Date:	Thu, 21 Oct 2021	**Prob (F-statistic):**	7.09e-63
Time:	13:30:40	**Log-Likelihood:**	32.143
No. Observations:	78	**AIC:**	-56.29
Df Residuals:	74	**BIC:**	-46.86
Df Model:	3		
Covariance Type:	nonrobust		

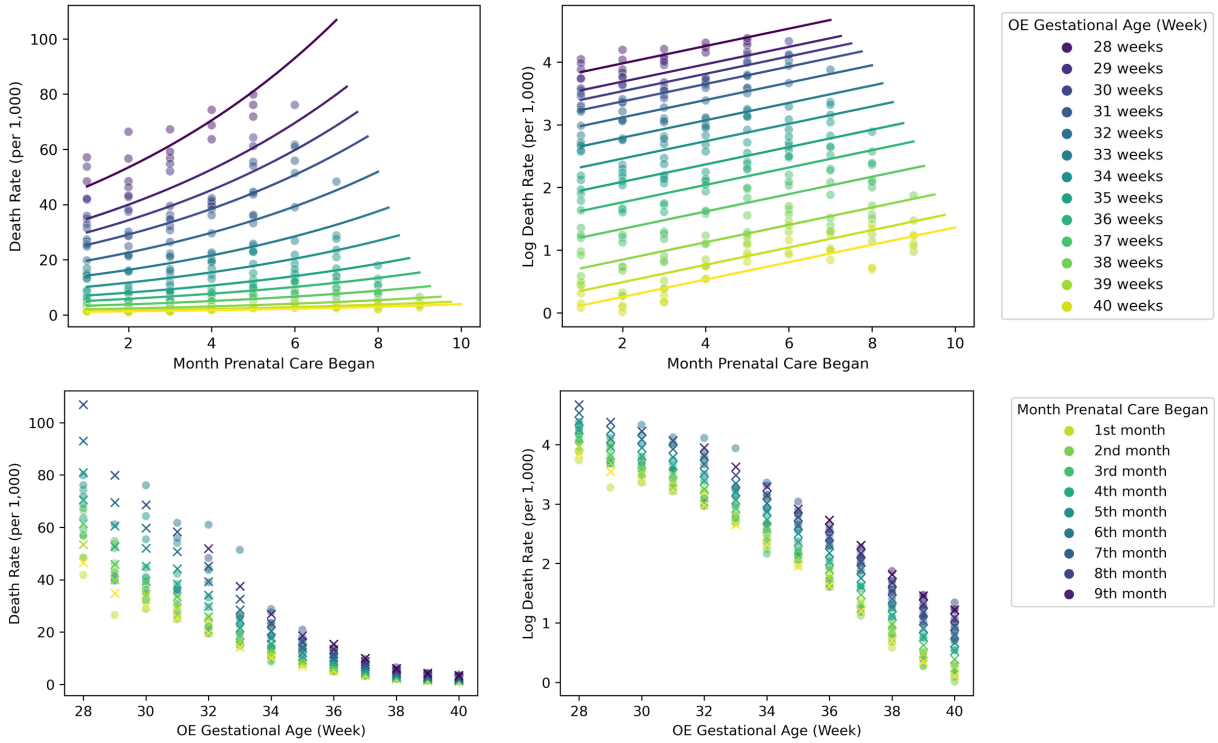
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8559	0.026	108.405	0.000	2.803	2.908
Week Centered	-0.2602	0.007	-36.961	0.000	-0.274	-0.246
Prenatal Binary	-0.4968	0.037	-13.335	0.000	-0.571	-0.423
Cross Term	-0.0632	0.010	-6.351	0.000	-0.083	-0.043

11 Gestational age in weeks
Above: Log-linear model to predict infant death rate per 1,000 live births based on prenatal care and gestational age in weeks. Models and data from 2016-2018 plotted in both linear (left) and log scales (right). Left: Model summary results.

DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

Next, we wondered if it mattered not only *if* a mother received prenatal care, but *when* she started care. Focusing on mothers who had prenatal care, we first fit a log-linear model on death rate based on month that prenatal care began and gestational age. Because this model overestimated death rates at the extremes of the gestational age range but underestimated it in between, we then fit a linear mixed-effects model on log death rate with a fixed effect on the month that prenatal care began and random intercepts on gestational age. While the statistically-significant coefficient for month prenatal care began was similar in both models, the mixed-effects model did a better job at fitting the data by gestational age. Overall this model confirms that beginning prenatal care earlier in a pregnancy is associated with decreased death rates, with a greater effect for infants born at earlier gestation.



Model:	MixedLM	Dependent Variable:	y
No. Observations:	266	Method:	REML
No. Groups:	13	Scale:	0.0337
Min. group size:	13	Log-Likelihood:	14.4111
Max. group size:	27	Converged:	Yes
Mean group size:	20.5		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Month Centered	0.139	0.005	25.923	0.000	0.128	0.149
Week Centered	8.099	17.705				

12 Month prenatal care began

Above: Linear mixed-effects model of log death rate with a fixed effect on month that prenatal care began and random intercepts on gestational age. Models and data from 2016-2018 plotted in both linear (left) and log scales (right), with color by gestational age (top) and month prenatal care began (bottom). Predictions are plotted as both lines (top) and x's (bottom). Left: Model summary results.

DATA ANALYSIS & COMPUTATION

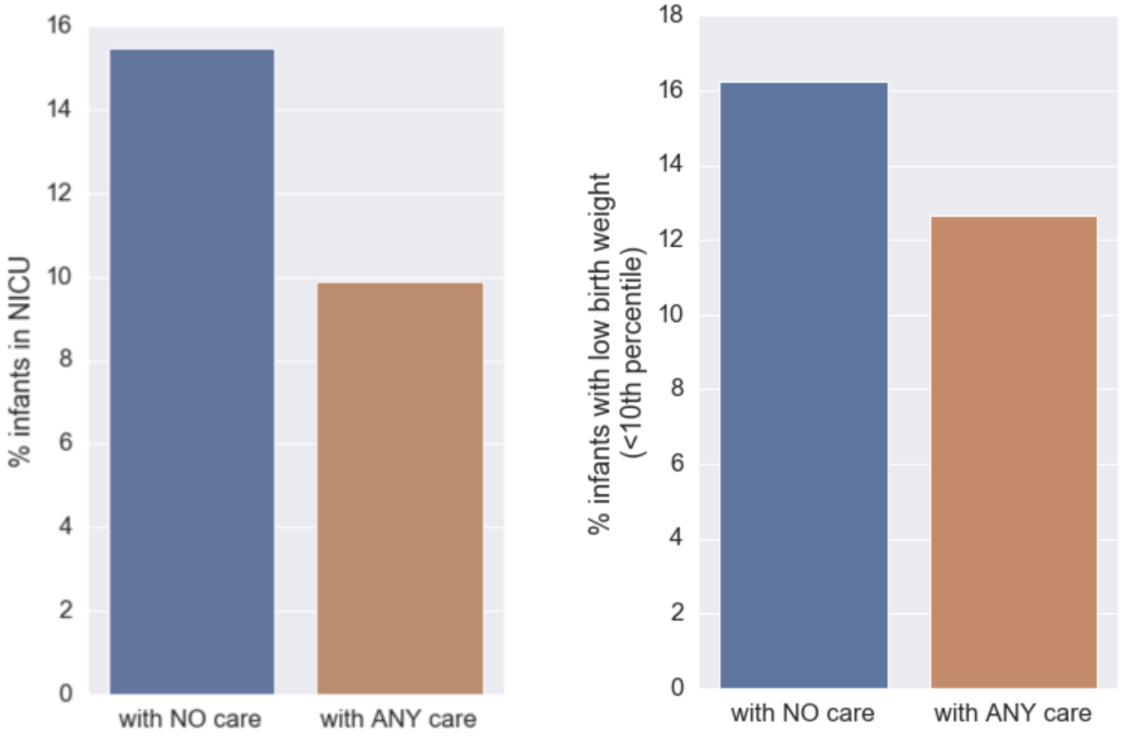
STATISTICAL DATA ANALYSIS

How does prenatal care affect birth outcomes?
PART II: NICU admission & birth weight

METHOD: Cases were matched on maternal features (age, education, birth place [within vs. outside US], race/ethnicity, use of WIC (supplemental nutrition program for low-income women), cigarette use before and during pregnancy, pre-existing medical conditions [diabetes, hypertension, or STD infection] as well as gestational age (in weeks) at the time of infant birth, delivery type (vaginal or cesarean), and the method of payment for birth expenses.*

RESULTS: In 2018, having access to prenatal care was associated with

- a 36% reduction in rate of NICU admissions, and
- a 22% reduction in very low birth weight (adjusted for gestational age).



13 NICU admissions and low birth weight
Comparing NICU admission rates and low birth weight rates with and without prenatal care after matching samples on selected maternal features, use of WIC, cigarette use, pre-existing medical conditions, gestational age, delivery type, and payment source for delivery.

*Cases limited to singleton births of first children occurring at or after 28 weeks gestation in 2018.

DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

Following up on the results of our matching analysis, we conducted a **logistic regression** to estimate **the effect of prenatal care on the odds of an infant entering the NICU after birth**, while holding key demographic, health, and birth factors constant.

In addition to binary-coded prenatal care, we included the following predictors in the model: maternal age, education, nativity [within vs. outside US], race/ethnicity, cigarette use during pregnancy, as well as gestational age (in weeks) at the time of infant birth, delivery type (vaginal or cesarean), and the method of payment for birth expenses.* Variables that did not significantly predict NICU admission were removed from the final model.

Consistent with the matching analysis, results of our logistic regression indicate that access to prenatal care ***decreased the odds of NICU admission by ~30%.***

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.6059	0.027	-94.922	0.000	-2.660	-2.552
Prenatal care	-0.3546	0.027	-13.228	0.000	-0.407	-0.302
Gestational week	-0.3603	0.002	-234.940	0.000	-0.363	-0.357
Maternal age	0.0596	0.004	15.337	0.000	0.052	0.067
Maternal education	-0.0171	0.003	-5.803	0.000	-0.023	-0.011
Mat. born outside US	0.0357	0.010	3.663	0.000	0.017	0.055
Race/ethnicity: black	0.0640	0.011	5.638	0.000	0.042	0.086
Race/ethnicity: hispanic	0.0424	0.010	4.223	0.000	0.023	0.062
Smoking during pregn.	0.1429	0.018	8.082	0.000	0.108	0.178
Cesarean section	0.7121	0.008	92.769	0.000	0.697	0.727
Payment: medicaid	0.0969	0.010	10.198	0.000	0.078	0.116
Payment: self-pay	-0.1154	0.022	-5.211	0.000	-0.159	-0.072

*Cases limited to singleton births of first children occurring at or after 28 weeks gestation in 2018.

DATA ANALYSIS & COMPUTATION

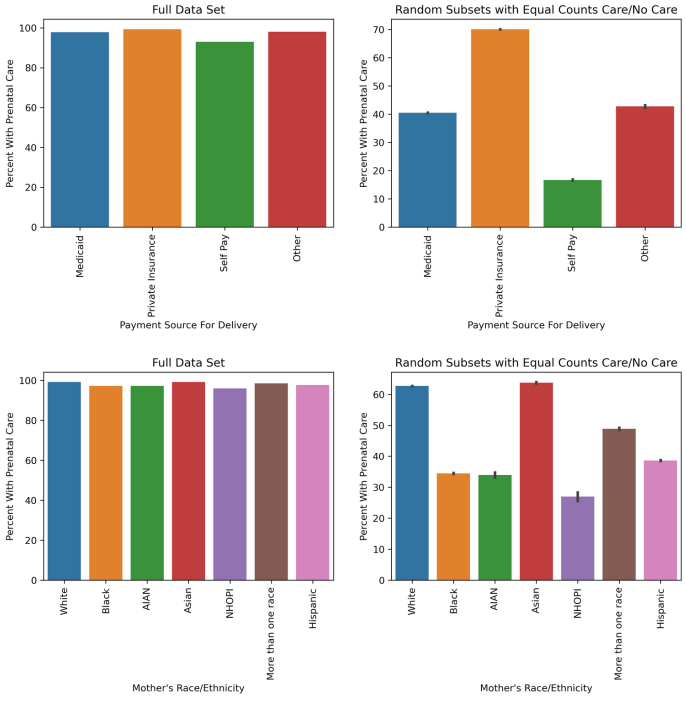
STATISTICAL DATA ANALYSIS

Question 2: What factors influence who receives prenatal care?

Because prenatal care is so important for healthy birth outcomes, we were interested in discovering which indicators predict the use of prenatal care, which could in turn inform outreach efforts for underserved populations. We considered 6 indicators in the 2018 Nativity Records data set:

1. Mother's age (categorical: 5-year bins from < 15 to 54 years)
2. Mother's nativity (binary: born inside or outside the U.S.)
3. Mother's race/ethnicity (categorical: 6 race categories plus Hispanic origin)
4. Mother's education (categorical: 8 levels from <= 8th grade to doctorate)
5. WIC (binary: participated in supplemental nutrition program)
6. Payment source for delivery (categorical: 4 payment methods)

To predict whether or not someone would receive prenatal care, we used logistic regression. Our initial model had 98.51% accuracy with 100% sensitivity and 0% specificity. Due to the fact that only 1.48% of the records had no prenatal care, the model was predicting that *everyone* received care. Because not receiving any prenatal care was rare, our next approach was to re-fit the model using random subsets of our data with equal counts of who received care and who did not. In addition to producing a more informative model, using balanced subsets of our data also resulted in more variability in who received care within most of our indicators (examples below).



14 Prenatal care by indicator
Left: Percent of records with prenatal care by payment source for deliver (top) and mother's race/ethnicity (bottom). Percentages are reported for both the full data set (left) and for 25 random subsets with equal counts for care and no care (right).

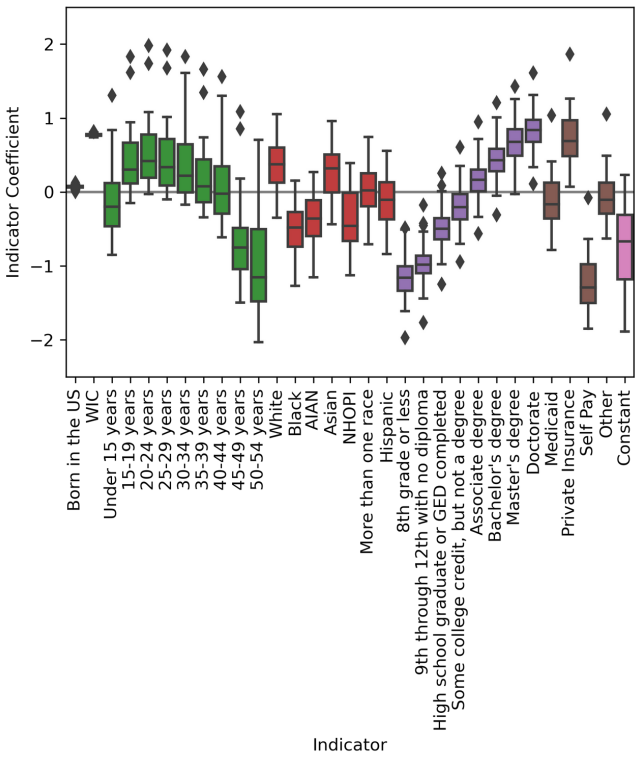
DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

On the right, we plot the distribution of indicator coefficients from our logistic regression models for the 25 different random subsets of our data (no interaction terms were included in the model). These 25 models had an average accuracy of 69.14% with 69.88% specificity and 68.37% sensitivity, so we sacrificed overall accuracy for the ability to detect records without prenatal care.

The sign of the coefficients can be interpreted in terms of how an indicator affects the log odds of a person receiving prenatal care. For example, the private insurance variable has a positive coefficient, so having private insurance is associated with an increase in the log odds of receiving care. On the other hand, the self pay variable has a negative coefficient, which can be interpreted to mean that not having insurance is associated with a decease in the log odds of receiving care.

Additionally, the magnitude of a coefficient can reflect the relative importance of the indicator. For instance, the coefficient for "being born in the U.S." is small relative to the others, indicating that this variable is less influential in predicting the log odds of receiving care.



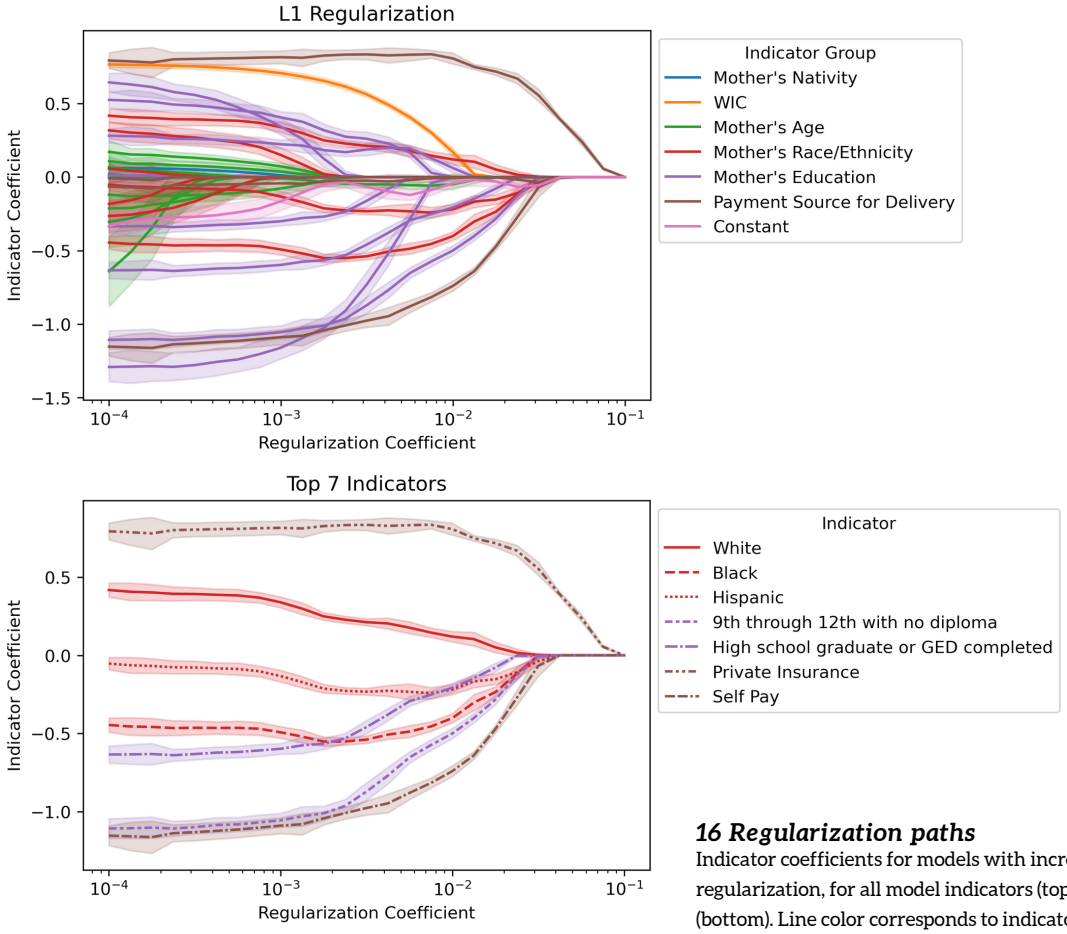
15 Indicator coefficients
Indicator coefficients for logistic regression models to predict who receives prenatal care, fit using 25 random data subsets with equal counts of who received care and who did not. Box colors correspond to indicator groups.

To further determine which of these indicators are most influential in predicting who received prenatal care, we added a L1-regularization term to our model. The L1-norm promotes sparsity in the solution vector, so as we increase the regularization coefficient, more of the indicator coefficients in the logistic function become zero. One interpretation of this trend is that the coefficients that remain non-zero for larger regularization coefficients are more important to the accuracy of the model. Furthermore, while the L1-norm knocks out *individual* coefficients as regularization increases, a group sparsity term (the sum of the L2-norm of groups of coefficients) can knock out *groups* of coefficients. These two approaches can be used to identify which individual indicators (e.g., private insurance vs. self pay) and which groups of indicators (e.g., payment source for delivery vs. mother's age) are more influential, respectively. Using our 25 data subsets, we fit models for increasing levels of regularization with both a L1-regularization term and a group sparsity term. Because both methods identified the same indicators, we limit our discussion to the results of the L1 regularization.

DATA ANALYSIS & COMPUTATION

STATISTICAL DATA ANALYSIS

Below, we see the regularization paths for our model coefficients, with coefficients for mother's nativity and age dropping out first and coefficients related to payment source for delivery, mother's race/ethnicity, and mother's education remaining for larger levels of regularization (the coefficient for WIC is not far behind). These results suggest that the latter indicators were the most important in our model, and could possibly be associated with populations that are less likely to receive prenatal care. Specifically, the most important indicators were whether or not payment source for delivery was private insurance or self pay, whether or not the mother was Hispanic, Black, or White, and whether or not the mother had attended high school, with or without a diploma or GED. Within these top indicators, our results suggest that Black and Hispanic mothers, and mothers without a college education, may be less likely to receive prenatal care, and could benefit from targeted outreach or expanded access to care.



16 Regularization paths
Indicator coefficients for models with increasing levels of regularization, for all model indicators (top) and the top 7 indicators (bottom). Line color corresponds to indicator groups.

CONCLUSIONS & FUTURE WORK

No. 01 — Model Results

Through a combination of log-linear modeling, linear mixed-effect modeling, matching, and regularized logistic regression, we observed that those with access to prenatal care had lower infant death rates, lower rates of admission to the neonatal intensive care unit (NICU), and a lower percentage of infants with very low birth weight. Furthermore, starting prenatal care earlier in the pregnancy was a significant factor in lowering death rates, with a greater impact for infants born at earlier gestational age. Finally, we found that differences in insurance status, race, and education were important indicators for who receives prenatal care. These results suggest that expanded access to and use of prenatal care, especially within specific population groups, could help to lower rates of infant morbidity and mortality in the U.S.

No. 02 — Model Improvements

There are a variety of modifications we could use to improve our models. For example, we learned the importance of matching and balancing data sets, which could be implemented in all of our models. Additionally, we could explore other variable transformations and encodings (e.g., treating educational attainment as a continuous rather than categorical variable when looking at who receives prenatal care). To achieve higher prediction accuracy, we could also include additional features and use more sophisticated models that capture the interactions between different indicators. Finally, there were many dimensions we did not get a chance to investigate, including spatial and temporal patterns.

REFERENCES

Cohen, Joshua. "U.S. Maternal And Infant Mortality: More Signs Of Public Health Neglect." Forbes, 1 Aug. 2021, www.forbes.com/sites/joshuacohen/2021/08/01/us-maternal-and-infant-mortality-more-signs-of-public-health-neglect

Martin, M.P.H., Joyce A., et al. "Births: Final Data for 2018." National Vital Statistics Reports, vol. 68, no. 13, Nov. 2019, p. 47.

Singh, Gopal K. "Trends and Social Inequalities in Maternal Mortality in the United States, 1969-2018." International Journal of Maternal and Child Health and AI DS, vol. 10, no. 1, 2021, pp. 29–42.

Data Sets

United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics (DVS). Linked Birth / Infant Death Records 2007-2018, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, on CDC WONDER On-line Database. Accessed at http://wonder.cdc.gov/lbd-current.html on Oct 22, 2021 5:10:08 PM

United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics, Natality public-use data 2016-2019, on CDC WONDER Online Database, October 2020. Accessed at http://wonder.cdc.gov/natality-expanded-current.html on Oct 22, 2021 5:08:55 PM

Code

- Prenatal Care Dashboard
- Infant Death Dashboard
- Predicting death rate by prenatal care and gestational age
- Predicting prenatal care by maternal indicators