Final project report:
Modeling longitudinal opioid overdose deaths in Virginia counties using an opioid risk-environment framework

Faysal Shaikh

CSI 672 - Fall 2021

The United States "opioid epidemic" is a widespread public health issue. With opioid overdose related deaths generally increasing since the introduction and heavy adoption of opioid drugs in the late 1900s and early 2000s in the United States, we are interested in understanding potential driving factors for overdose deaths in our geographic location of Virginia. We perform linear mixed-effects analyses to understand the problem from the perspective of social, health, and other factors as described in an "opioid risk-environment framework." We explore our results and provide recommendation for further analyses.

Modeling longitudinal opioid overdose deaths in Virginia counties using an opioid risk-environment framework

Faysal Shaikh **CSI 672 - Fall 2021**

Table of contents

List of figures

List of tables

1 Introduction

The purpose of this project is to perform statistical analyses on longitudinal data relevant to the United States "opioid epidemic." We are interested in understanding the potential driving factors for this widespread public health issue within the counties of Virginia. This project is in part a collaboration with the Justice Community Opioid Innovation Network (JCOIN) and is thus also inspired by work of the National Institutes of Health (NIH) Helping End Addiction Long-term (HEAL) Initiative (NIH HEAL Initiative, 2021). We begin our exposition by describing the United States "opioid epidemic" in the text below.

1.1 Origins: The undertreatment of pain and rise of opioid drugs

Prior to the 1980s, opioid drugs were not considered commonplace treatments for chronic pain in the United States. There is actually considerable evidence of a characteristic "opiopohobia," especially following the 1914 Harrison Narcotic Tax Act (Jones et al., 2018). The addictive potential of opioid drugs was understood following the American Civil War, during which these drugs were used in field hospitals to relieve surgery pains (Provine, 2011).

As such, this time period saw opioid addiction generally viewed through a medical lens. Unbeknownst to many, buildup to the passage of the Harrison Act potentially served as early warning signs of the impending nefarious "War on Drugs." In fact, President Theodore Roosevelt's appointed Opium Commissioner in 1908 had used explicitly racial claims, blaming opium for illicit sexual relations between white women and Chinese men and blaming cocaine for violence in African American men, to push for drug control at the federal level (Provine, 2011). We now know that similar techniques were used by Henry Anslinger, who from 1930 until 1962 served as the first commissioner of the Federal Bureau of Narcotics, to incite racial fearmongering and reshape the general view of addiction towards one of criminality via what we call today the "War on Drugs" (Provine, 2011).

Views of opioid medications began to change during the 1980s, as literature emerged highlighting the undertreatment of pain in the United States. These findings coincided with the surfacing of two pieces of literature, neither of which are considered to meet today's standard for scientific rigor, regarding an apparent low addiction potential for opioid drugs (Jones et al., 2018). Prior to this time, opioid prescriptions were typically reserved for short-term pain relief following surgery or cancer patients suffering from terminal illness. However, a burgeoning interest in the utility of opioid drugs for non-cancer pain, at times driven by misconceptions of non-cancer pain by underinformed cancer pain specialists, began to take hold of the medical community (Jones et al., 2018) and would lead to a gradual increase of opioid prescriptions during this time period (DeWeerdt, 2019).

The following years saw many notable patient advocacy and regulatory organizations, such as the American Pain Society (launching their "pain as a fifth vital sign" campaign in 1995), the Veteran's Health Administration (moving to adopt "pain as a fifth vital sign" in 1999), the Joint Commission (publishing standards for pain management in 2000), the Institute of Medicine, the Federation of State Medical Boards, and even the United States Drug Enforcement Agency, synergistically pushed for a more structured approach to pain assessment and management that heavily relied upon the prescription of opioid drugs (Jones et al., 2018). It was also this time period that saw pharmaceutical companies devote significant resources towards lobbying, sponsorships, and marketing to promote their opioid products to the greatest extent possible (DeWeerdt, 2019).

Some pharmaceutical companies took advantage of these opportunistic circumstances by peddling fraudulent claims to sell their opioid products. Purdue Pharma falsely marketed OxyContin, a new sustained-release formulation of the highly-addictive opioid drug oxycodone, as less addictive than other opioid painkillers; Purdue Pharma later admitted their knowledge of OxyContin as addictive in a 2007 lawsuit (DeWeerdt, 2019). Purdue even focused their initial marketing of OxyContin towards white communities, knowing that the image of the typical drug addict painted by Anslinger and the ensuing War on Drugs would serve their message of OxyContin as a non-addictive drug (DeWeerdt, 2019). As a result of these efforts, OxyContin prescriptions rose sharply from 670,000 in 1970 to 6.2 million in 2002 (Jones et al., 2018).

Despite best efforts from the pharmaceutical industry to hide the truth about the addictive potential of opioid drugs, it was only a matter of time before the reality of the situation would reveal itself.

1.2 Today: The opioid overdose crisis

The modern opioid epidemic in the United States is often described as taking place in 3 overlapping phases (DeWeerdt, 2019). The first phase began with the overprescription and abuse of opioid pharmaceuticals described earlier. The second phase, heavily involving heroin, saw heroin overdose deaths increase nearly fivefold in the United States from 2010 to 2016 (DeWeerdt, 2019). The early days of the third (present) phase saw the involvement of cheaper yet more potent opioids, namely fentanyl, such that opioid deaths from fentanyl and similar molecules increased by 88% per year between 2013 and 2016 (DeWeerdt, 2019).

American opioid overdose deaths in 2016 surpassed 42,000, at that point in time more than any previous year on record (U.S. Department of Health and Human Services, 2021). This record was subsequently broken by over 47,000 opioid overdose deaths in 2017 (National Institute on Drug Abuse, 2021). Despite the declaration of the opioid epidemic as a public health emergency in 2017 (U.S. Department of Health and Human Services, 2021), which saw a decline in opioid overdose deaths from 2017 to 2018, the previous record was surpassed once again by nearly 50,000 opioid overdose deaths in 2019 (National Institute on Drug Abuse, 2021).

Although pandemic coronavirus disease 2019 (COVID-19) has resulted in challenges obtaining recent unbiased estimates of opioid use (Haley & Saitz, 2020), we can safely continue to assume the omnipresence of the opioid epidemic in the United States today.

2 Theory

We begin by describing so-called "social determinants of health" and move towards a more comprehensive discussion of the notable "risk environment framework" (Rhodes, 2002) that has started to see successful adoption in recent works (M.A. Kolak et al., 2020).

2.1 Social determinants of health (SDoH)

As its name suggests, the social determinants of health (SDoH) framework provides a valuable conceptualization for how various social and economic factors may play a deterministic role in health outcomes. Much early work inspiring the SDoH has shown associations between socioeconomic factors and health outcomes, but the specific causal nature behind the health contributions of specific social factors may vary (Wilkinson et al., 2003).

For example, the social factor of race may play a role in adverse health outcomes through both direct and indirect mechanisms of systemic racism and discrimination. Medical students with inaccurate perceptions of black patients as having "thicker skin" and higher pain tolerance than white patients are shown to minimize their black patients' pain presentation and offer accordingly poorer treatment recommendations (Hoffman et al., 2016). Black mothers typically face higher levels of low birth weight and preterm-birth than the general population, but a dedicated midwifery program was shown to numerically rectify the disparity entirely (Josephs & Brown, 2017). The infamous Tuskegee syphilis study (Brandt, 1978), which violated several ethical considerations but namely refused to provide adequate medical treatment (sometimes even involving deceiving the participants, e.g., convincing them that they were receiving medical treatment when they were not) to the black study participants despite the presence of widely-available and successful medications for syphilis, has tainted the image of the biomedical research enterprise to the point that the black community in the United States have a warranted level of distrust in the medical system; this may potentially lead to patients from the black community prolonging their illness prior to seeking help, which could potentially exacerbate issues that otherwise would have been minimized with appropriate early-intervention or preventative care.

This above is only a small amount of consideration given to one category of one socioeconomic factor, yet the potential importance of considering socioeconomic factors in health research is relatively apparent.

2.2 Risk environment frameworks

While the SDoH framework provides one additional dimension to the view of health as more than simply health outcomes, the risk environment framework originally proposed by Rhodes (2002) takes this a step further by shifting perspective from the health and socioeconomic factors of an individual towards also understanding the health-influencing factors may be present in their relative environment.

In particular, the original work of Rhodes (2002) attempted to conceptualize the environmental factors that may influence the propensity for drug-related harm, such as HIV infection associated with drug injection, and to better prepare public health practitioners to practice effective harm-reduction in the communities they serve.

The risk-environment framework has been utilized effectively relatively recently in the work of M.A. Kolak et al. (2020) to understand opioid-related overdose in rural Southern Illinois. The authors of this work utilize the risk environment framework to describe how rural environments can exhibit uniquely vulnerable characteristics that may culminate in risk "hotspots" of high overdose rates and other adverse health outcomes (M.A. Kolak et al., 2020). The authors utilize variables that reflect several spheres of influence: social, economic, policy, and physical (M.A. Kolak et al., 2020). With an understanding of the historical context of the opioid epidemic in the United States, the inclusion of these additional spheres of influence has been shown to better inform researchers' view of the problem.

3 Research hypothesis

Our approach utilizing a risk-environment framework to study opioid overdose deaths for counties in Virginia will provide us with inferences of the driving factors of the "opioid epidemic." In particular, our statistical hypotheses take the form of inferences on the parameters of our regression model.

A general hypothesis we may have about each given factor in consideration is whether that factor is a statistically significant predictor in our regression model. In other words, we would like to perform a hypothesis test against the null hypothesis that the coefficient of a given predictor in our regression is 0. We would consider a rejection of this null hypothesis as a statement that a given predictor is a driver of the response variable, within the boundaries of our dataset.

It should be noted that inference techniques utilized in the standard ordinary least squares (OLS) regression approach, in which model parameters converge towards "nice" asymptotic distributions, do not necessarily transfer for the more sophisticated "mixed-effects model" (described later) often used in this work (University of Wisconsin-Madison Social Science Computing Cooperative, 2016).

We may also be interested in a hypothesis test against a null hypothesis that the variance of random effects (random intercepts per county to account for data heterogeneity between counties) is 0. We would consider a rejection of this null hypothesis as a statement that a special consideration of the data (as given by the prespecified random effects structure) is required to appropriately model the data.

As described above, special consideration must be taken in performing these tests to ensure that they are handled properly in the context of mixed models (University of Wisconsin-Madison Social Science Computing Cooperative, 2016).

4 Experiment design

We begin our exposition by describing notable previous work in our area of interest, and follow this by describing our proposed approach and tools.

4.1 Previous work

We begin with the work of Heyman et al. (2019), providing evidence for the importance of socioeconomic factors in contributing to drug overdose deaths at the state level. This study identifies several correlates for drug overdose deaths, including availability of opioid prescriptions, percent in the labor force, elementary school English and math national test scores, teen birth rate, and many more. Notably, the researchers found the availability of opioid prescriptions in a given state to be more valuable as a predictor when in a regression model for a response stratified by the "non-hispanic white" race category than for the same response stratified by the "hispanic/non-white" (i.e., "minority") race category (Heyman et al., 2019). The study additionally utilizes "random and fixed effects panel regressions" as their statistical method of choice.

We now turn to the work of Kim and Yang (2020), highlighting county-level analyses of opioid-related overdose deaths. This work highlights demographic group variables, socioeconomic group variables, and insurance status as predictors for the response variables of all-opioid overdose deaths, heroin overdose deaths, and fentanyl overdose deaths. This work also utilizes the multilevel mixed-effects models approach. Their findings seem to also illustrate the utility of using a model containing socioeconomic variables in explaining opioid overdose deaths. Notably, this work does not account for opioid prescription practices in the analyses, but mentions a desire for further in-depth analyses as on of the conclusions.

The work of Bauer et al. (2021) analyzes the longitudinal opioid-suspected overdose in the Houston metropolitan area. This work represents a departure from the previously-described works in methodology, as the authors of this work utilize a "Bayesian spatiotemporal modeling" framework rather than that of the typical frequentist mixed-effects model approach for longitudinal data. An advantage of this approach is that it potentially handles spatial relationships in a more appropriate manner than assigning random intercepts to different geographic areas. The authors of this work find that their various considered socioeconomic variables (but no variables directly related to opioid prescriptions) seem to be associated with differences in opioid overdose deaths at the zipcode-level in the Houston metropolitan area via 3 different spatial modeling approaches.

4.2 Our approach

We decide to structure the design of our experiments based on the successes of previous work, as well as data considerations and researcher expertise.

Our statistical analyses were conducted via linear mixed-effects models allowing for random intercepts by county to account for heterogeneity of data observed in different counties. We restrict our analyses in a single state, in this case Virginia, to deal with both missing data issues (minimizing our sample) and to minimize the effects of different political regulation (i.e., different state government entities) between entities. Although counties certainly have different governing entities, we hope that our implementation of random intercept may also numerically account for this (assumed to be) relatively minor differences when compared with the potentially more apparent effects of different state governments.

Regression predictors were selected based on the findings of previously-mentioned studies. Previous studies showcased a heavy focus on socioeconomic status (SES) factors. As such, we plan to include socioeconomic factors in our work, but depart from a sole focus on socioeconomic factors to adopt a more complete "risk environment framework" view that includes health factors and other non-SES variables. As teen birth rate was explicitly mentioned in many of our studies, we attempt to utilize longitudinal teen birth rate as a time-varying covariate on a per-county basis. Although one of the three aforementioned studies included the availability of opioid prescriptions as a relevant factor in prediction (Heyman et al., 2019), the finding that this predictor was differentially important in understanding the overdose deaths in different race cateories, we opt to not include a variable of this nature in our models for aggregated county opioid deaths.

5 Data

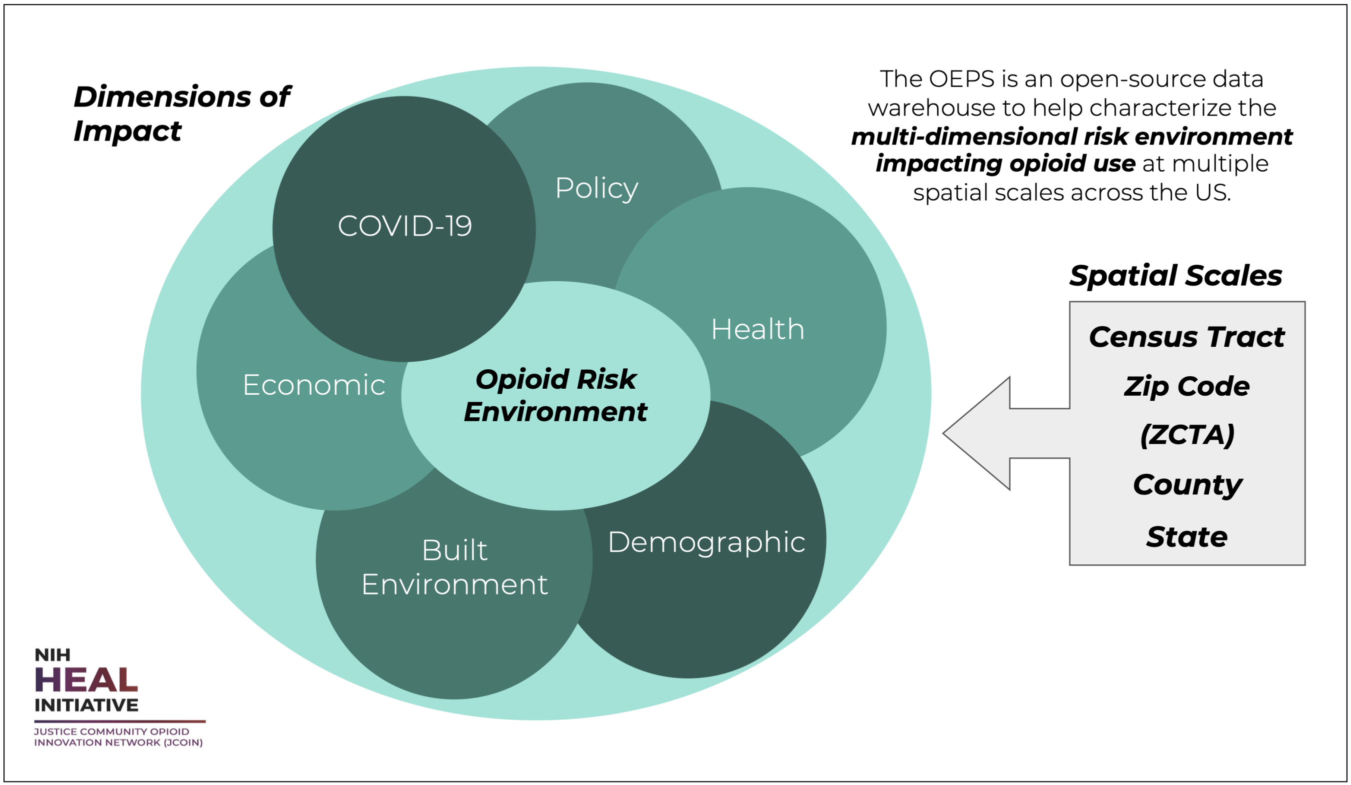
While numbers of opioid overdose deaths and other opioid-use-specific measures may provide useful information in understanding the scope of the opioid epidemic, they certainly do not provide the entire picture. Considering various social determinants of health (SDoH), the importance of which was especially validated in monitoring disease spread during the COVID-19 pandemic, may be required to better understand the factors that may contribute to opioid overdose outcomes.

Additionally, statistical relationships are necessarily derived at a group level. Thus we must ensure our perspective is not of individuals but of larger units of aggregation, e.g., geospatial neighborhoods. An additional benefit of considering these larger units is the ability to compare areas based on their differing profiles of SDoH and other environmental factors, for example by quantifying a multi-dimensional "risk environment" as described by Rhodes (2002).

As such, this work serves to highlight two distinct data sources which may prove relevant in understanding SDoH and additional environmental factor profiles of specific levels of geospatial aggregation: the University of Chicago Opioid Environmental Policy Scan (OEPS) dataset (Kolak et al., 2021), and the University of Wisconin County Health Rankings & Roadmaps (CHR) dataset (University of Wisconsin Population Health Institute & Robert Wood Johnson Foundation, n.d.).

5.1 The Opioid Environmental Policy Scan (OEPS) dataset

The OEPS dataset, developed by Kolak et al. (2021) as a collaboration between the University of Chicago Healthy Regions & Policies Lab and the University of Chicago Center for Spatial Data Science, utilizes a "risk environment framework" approach, as first described by Rhodes (2002), to consolidate data from various sources into six "spheres of influence": policy, health, demographic, economic, built environment, and COVID-19 (Kolak et al., 2021). This conceptual model for this data is shown in Figure 5.1 below (on page 8).

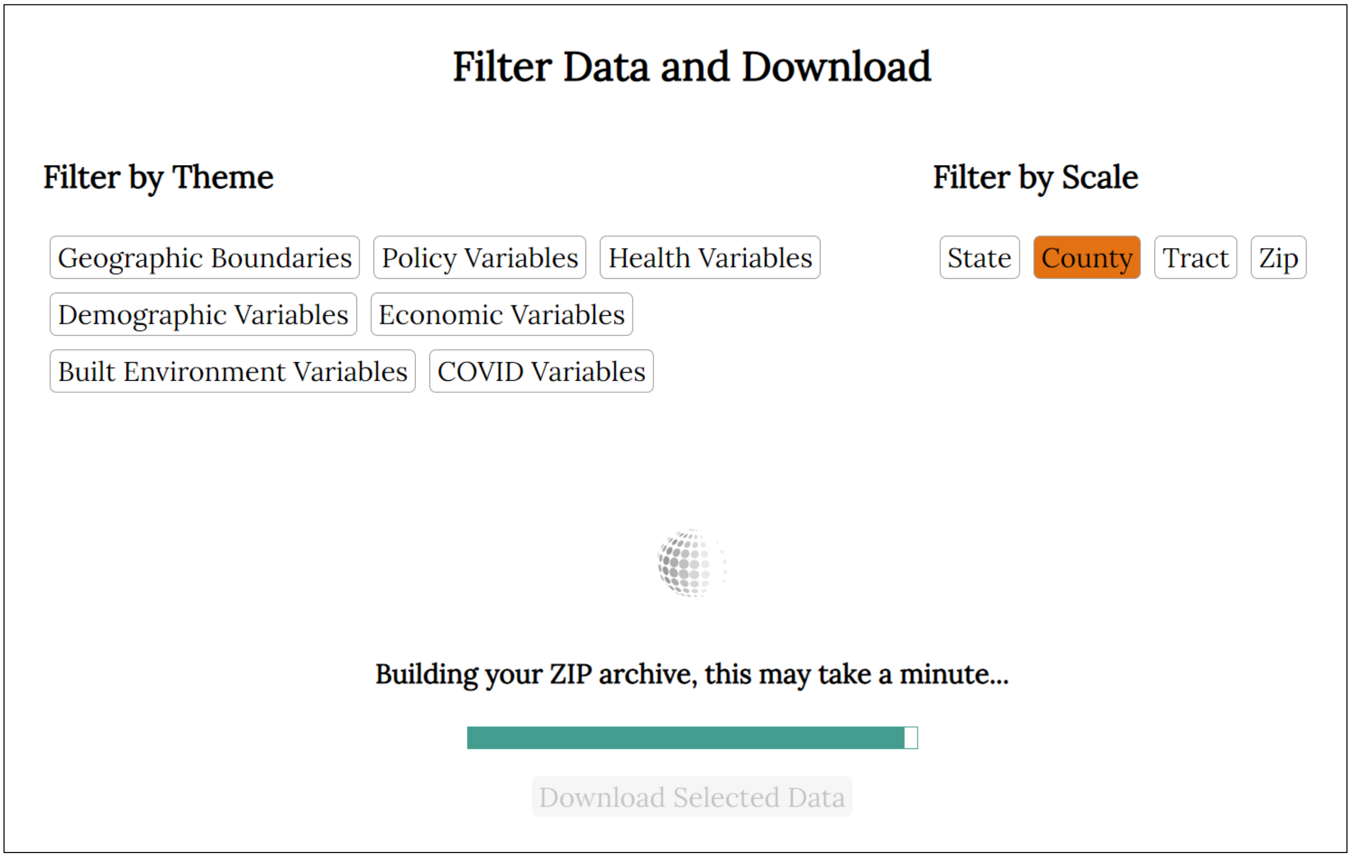


The OEPS is an open-source data warehouse to help characterize the ***multi-dimensional risk environment impacting opioid use*** at multiple spatial scales across the US.

Figure 5.1: Diagram of OEPS dataset.

The OEPS dataset is a nationwide extension of previous work by M.A. Kolak et al. (2020) highlighting the utility of the risk environment approach in understanding various health outcomes, including opioid-related overdose, in rural Southern Illinois between 2015 and 2017. In appreciation of “rural risk environment” analyses (M.A. Kolak et al., 2020) and of the expertise of the University of Chicago as the JCOIN Methodology and Advanced Analytics Resource Center (MAARC) (Kolak et al., 2021), this project hopes to utilize OEPS data to start with a multi-dimensional risk environment framework of the opioid epidemic at varying geospatial scales across the entire United States.

OEPS data were obtained via the “Filter Data and Download” section of the OEPS Explorer web application (https://oeps.netlify.app/download). “County” was selected under the “Filter by Scale” heading and downloaded to the project working directory. The data download interface is shown below in Figure 5.2 (on page 9).


Filter Data and Download

Filter by Theme

Filter by Scale

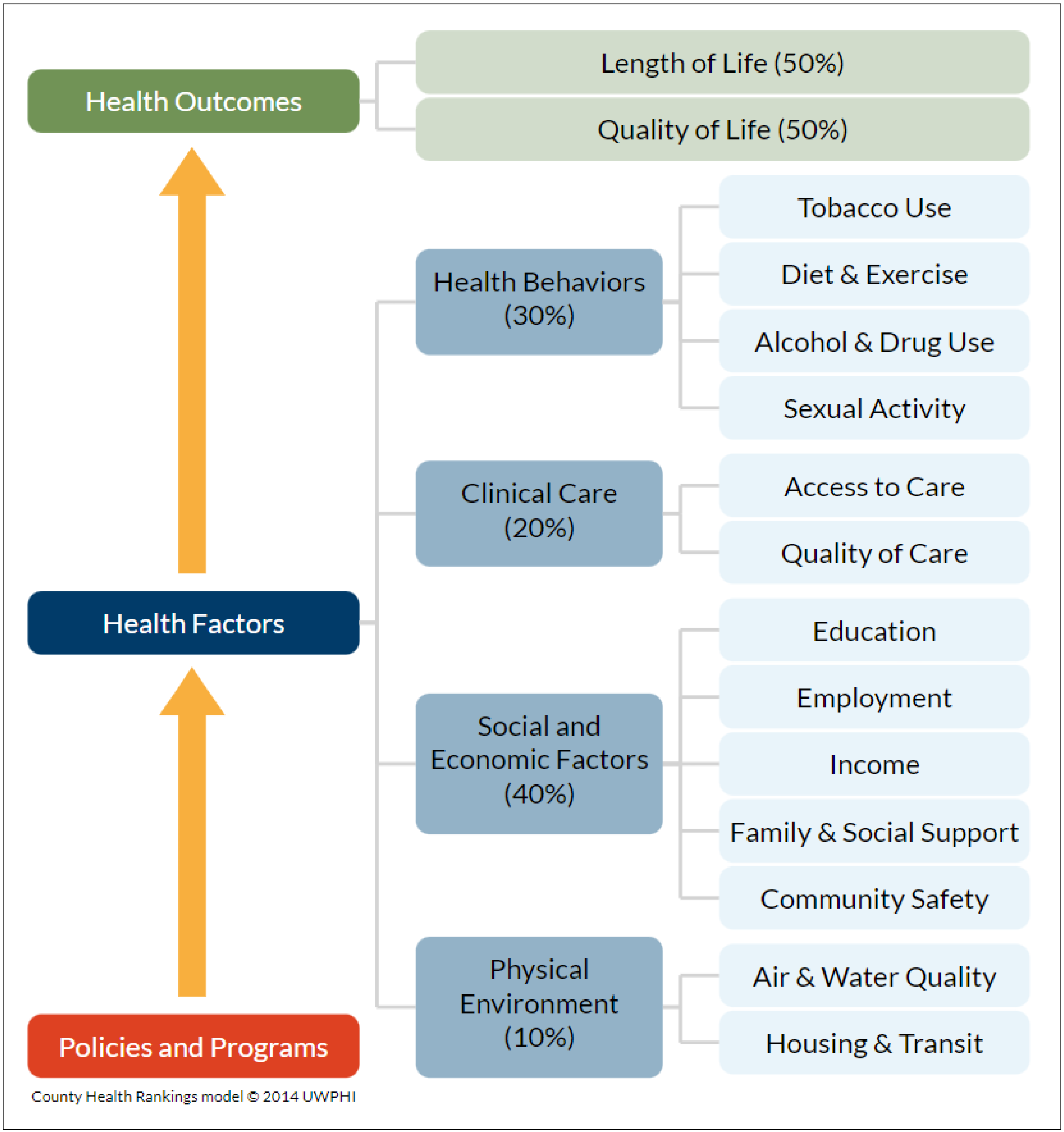
[✓ County]

Building your ZIP archive, this may take a minute...

Figure 5.2: OEPS Explorer data download interface.

5.2 The County Health Rankings & Roadmaps (CHR) dataset

The CHR dataset (University of Wisconsin Population Health Institute & Robert Wood Johnson Foundation, n.d.), released annually since 2010, is a collaborative effort between the University of Wisconsin Population Health Institute and the Robert Wood Johnson Foundation to consolidate data from various sources and publish health rankings that consider both health outcomes and modifiable health factors for each of over 3,000 counties and county equivalents in the United States (Remington et al., 2015). The conceptual model for CHR data is shown in Figure 5.3 (on page 10) below.



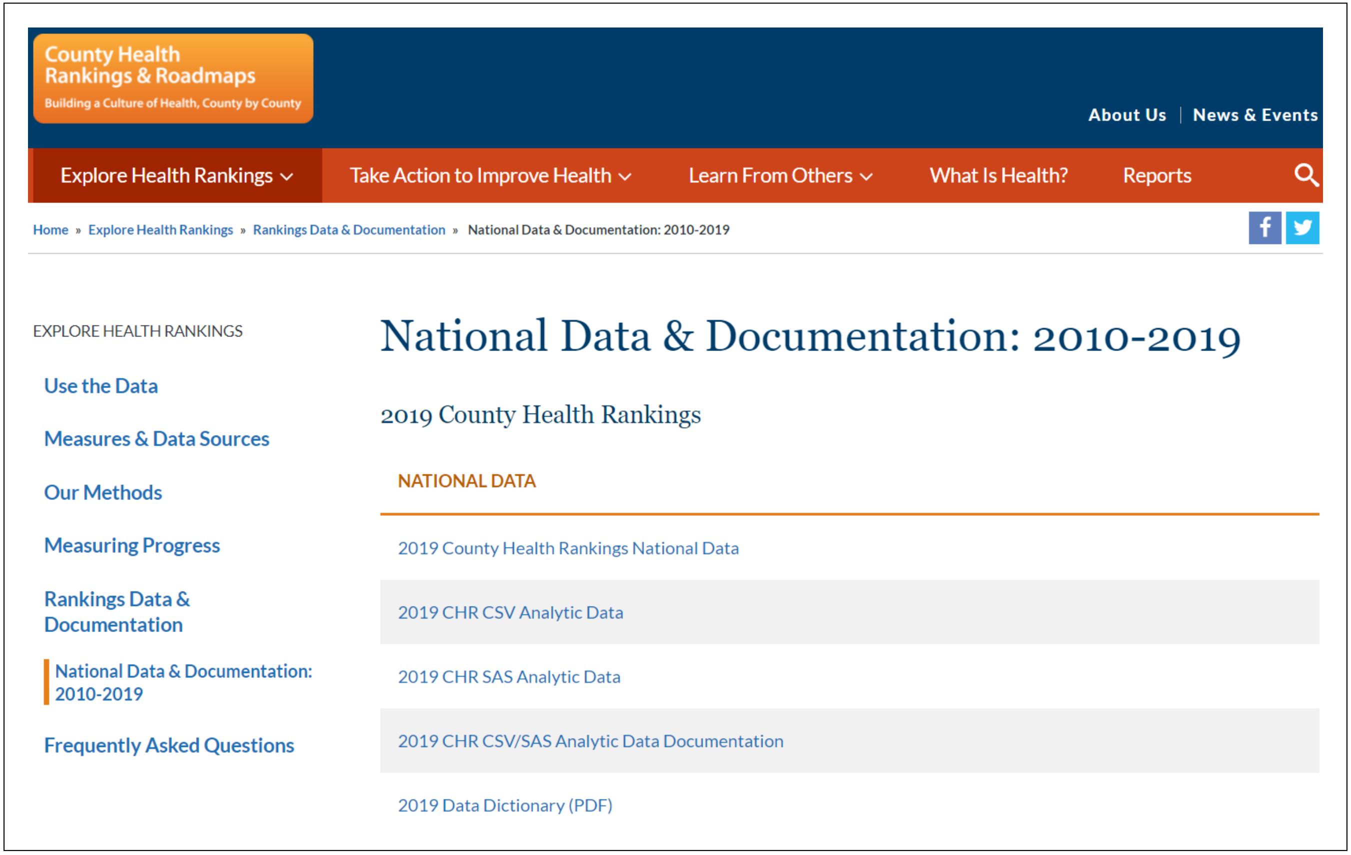
County Health Rankings model © 2014 UWPHI

Figure 5.3: Diagram of CHR dataset.

Although CHR data provide a more general picture of health and health factors when compared to the more opioid-use-related focus of OEPS data, the utility of these data in drawing conclusions of health within a county are well-demonstrated. For example, Remington et al. (2015) analyzed 2014 CHR data and found premature death rates to be more than twice as high in bottom five healthy counties when compared with top five healthy counties in each state.

While the dimensional richness of the multifaceted opioid risk environment is a strength of the OEPS dataset, data for each geospatial entity (without considering dis-/re- aggregation across multiple spatial scales) are only collected at a single timepoint, or "cross-sectionally" in time. As a result, face-value similiarites (or differences) seen "between subjects" (in this case, between geospatial entities, such as counties) may otherwise be characterized differently when taking into account a bigger picture that also includes "within-subject" variation over time (for example, repeated measurements of within the same geospatial entity), also known as "longitudinal" data. By nature of its yearly releases, CHR data are in fact longitudinal. This project thus attempts to consolidate cross-sectional and longitudinal data regarding SDoH and other environmental factors by merging the OEPS and CHR datasets in order to paint a more detailed multidimensional spatial and longitudinal picture of the opioid risk environment in the United States.

CHR data were obtained via either the "Rankings Data & Documentation" section (https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation) for 2020-2021 data, or via the "National Data & Documentation: 2010-2019" (https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019) section for historic data, previewed below in Figure 5.4 (on page 12). Relevant "County Health Rankings National Data" files were downloaded to the project working directory.


County Health Rankings & Roadmaps
Building a Culture of Health, County by County
About Us
News & Events
Explore Health Rankings
Take Action to Improve Health
Learn From Others
What Is Health?
Reports
Home » Explore Health Rankings » Rankings Data & Documentation » National Data & Documentation: 2010-2019
EXPLORE HEALTH RANKINGS
Use the Data
Measures & Data Sources
Our Methods
Measuring Progress
Rankings Data & Documentation
National Data & Documentation: 2010-2019
Frequently Asked Questions

National Data & Documentation: 2010-2019

2019 County Health Rankings

NATIONAL DATA

2019 County Health Rankings National Data

2019 CHR CSV Analytic Data

2019 CHR SAS Analytic Data

2019 CHR CSV/SAS Analytic Data Documentation

2019 Data Dictionary (PDF)

Figure 5.4: CHR historic data files download page.

5.3 Longitudinal opioid overdose outcomes

We conclude our discussion of data sources with the discussion of longitudinal opioid overdose outcomes data to serve as the response variable for our regression analyses.

As we are interested in obtaining reliable data regarding opioid overdose deaths at the County level, we turn to the United States CDC (Centers for Disease Control and Prevention) WONDER (Wide-ranging Online Data for Epidemiologic Research) database (Centers for Disease Control and Prevention, National Center for Health Statistics, 2020). Notably, we utilize a similar approach as that adopted by a previous work (Heyman et al., 2019) in terms of the specific cause-of-death codes specified in our query. However, we also ensure to query the data to produce a longitudinal outcome variable based on the available years of data in our longitudinal CHR dataset described above. Our complete query criteria were reproduced in Figure 5.5 (on page 13) below. (NOTE: Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.)

Query Criteria

States:	Virginia (51)

UCD - Drug/Alcohol Induced Causes:	Drug poisonings (overdose) Unintentional (X40-X44);
	Drug poisonings (overdose) Suicide (X60-X64);
	Drug poisonings (overdose) Homicide (X85);
	Drug poisonings (overdose) Undetermined (Y10-Y14)

Year/Month:	2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019
Group By:	County; Year
Show Totals:	Disabled
Show Zero Values:	True
Show Suppressed:	True
Calculate Rates Per:	100,000
Rate Options:	Default intercensal populations for years 2001-2009 (except Infant Age Groups)

Figure 5.5: CDC WONDER database query criteria: longitudinal opioid overdose deaths.

6 Data analysis

We begin our analyses with the descriptive statistics of our final merged dataset, followed by an overview of exploratory data visualization, and finally the results of our linear mixed-effects regression analyses.

6.1 Sample descriptive statistics

We begin our work by performing preprocessing on the relevant OEPS, CHR, and combined OEPS and CHR (also merged with CDC WONDER outcomes) datasets. The relevant preprocessing scripts for each of these steps are included in Appendix A.1.1 (on page 26), Appendix A.1.2 (on page 27), and Appendix A.1.3 (on page 29) for OEPS, CHR, and combined datasets, respectively.

Following the variable selection and merging of data highlighted in the relevant preprocessing scripts, we generated descriptive statistics tables for our sample via the initial lines of code found in the `regression-analyses.R` script in Appendix A.2.2 (on page 32). Our data sample is composed of both longitudinal and cross-sectional data, but we are interested in the unique cross-sectional predictors of our model and present the entire sample's cross-sectional descriptive statistics in Table T.1 (on page 14). Our value of N in the table represents the number of Virginia counties we have data for.

However, the descriptives table showcases a level of missingness in one of our variables: Alcohol density (per sq. mi.). We removed datapoints with missing values for this variable and recreated the descriptives table in Table T.2 (on page 14). Following removal of outliers, our new N Virginia counties of complete data is also reflected in the table.

	Overall cross-sectional sample (N=99)
Housing unit density (per sq. mi.)	
Mean (SD)	160 (500)
Alcohol store density (per sq. mi.)	
Mean (SD)	0.038 (0.13)
Missing	32 (32.3%)
Percentage rental units	
Mean (SD)	26 (8.2)
Percentage vacant units	
Mean (SD)	17 (9.4)
Percentage over 25 without HS diploma	
Mean (SD)	14 (5.2)
Percentage population disabled	
Mean (SD)	15 (4.5)
Percentage population white	
Mean (SD)	78 (16)

Table T.1: Cross-sectional descriptive statistics for the sample.

	Overall cross-sectional sample (N=67, after removing missing data)
Housing unit density (per sq. mi.)	
Mean (SD)	210 (600)
Alcohol store density (per sq. mi.)	
Mean (SD)	0.038 (0.13)
Percentage rental units	
Mean (SD)	26 (8.7)
Percentage vacant units	
Mean (SD)	16 (10)
Percentage over 25 without HS diploma	
Mean (SD)	13 (5.0)
Percentage population disabled	
Mean (SD)	15 (4.6)
Percentage population white	
Mean (SD)	79 (13)

Table T.2: Cross-sectional descriptive statistics for the sample without missing data.

As mentioned earlier, we focus on generating descriptives for our cross-sectional data, and shall dedicate effort to understanding the patterns in our longitudinal data via exploratory data visualization in the section below.

6.2 Exploratory data visualization

To try to understand our longitudinal data, we attempt to create trajectory plots (an intuitive extension of the concept of cross-sectional scatter plots with time prespecified on the horizontal axis) of our data. We begin by plotting our outcome variable for each Virginia county, opioid overdose related deaths, in Figure 6.1 (on page 15).

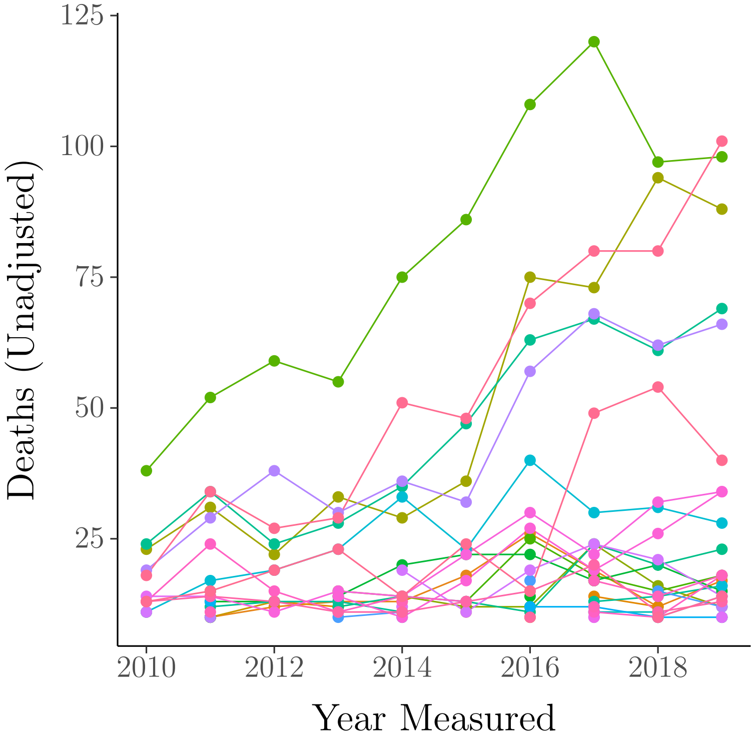


Figure 6.1: Opioid overdose death longitudinal trajectories for VA counties.

In the above plot we see different colors that represent each of the different Virginia counties that we have data for. Our outcome is taken from the CDC WONDER database, which takes into consideration data censoring procedures to preserve anonymity. Thus, we have some level of missingness in our data that is possible hidden by a floor effect seen with this specific measure (as no counties can have less than 0 deaths).

Additionally, it would not be instructive for us to regress simply on deaths without taking into account the population of each county. As such, in Figure 6.2 (on page 16) we visualize longitudinal trajectories of a crude death rate, calculated by dividing deaths in a county by the county population and multiplying by 1,000 to return a rate per 1,000 people. Although there is still a floor effect in this data, the data are more vertically spread, and thus the missingness becomes more visibly apparent.

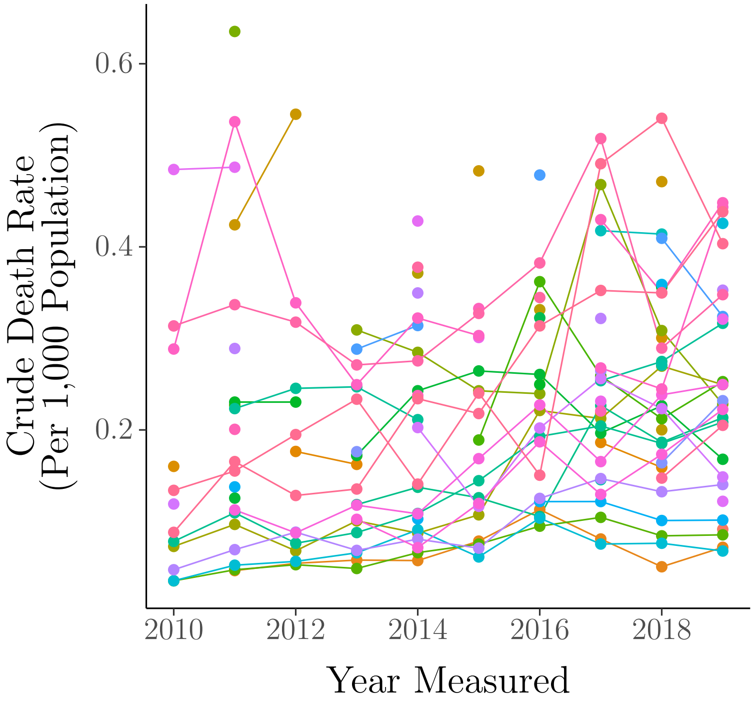


Figure 6.2: Opioid overdose crude death rate longitudinal trajectories for VA counties.

Our final longitudinal variable of consideration is a county's teen birth rate. This variable was utilized in multiple studies from previous work in this area, and we hope to see if our work replicates previous findings. This data is from the CHR dataset, which is an extremely well-funded and robust project that has extremely rich data. As expected, our visualization in 6.3 (on page 17) shows no missingness in this data.

As a final exploratory visualization, we are interested in exploring collinearity between our planned regression predictors. Predictor collinearity is an issue that can affect our results, so we attempt to understand this via a heatmap of correlation coefficients in Figure 6.4 (on page 17). From previous descriptions of the variables of our interest, some of these predictors take on percentage values, some take on rates (often per 1,000), and even others take on count-type values. As such, we utilize the Spearman rank correlation coefficient as a nonparametric alternative to the standard Pearson product-moment correlation coefficient. The heatmap shows that most variables have low correlation, however, we may want to be careful with building models containing both alcohol store density (alcDens) and housing unit density (unitDens) as these variables exhibit Spearman's $\rho = 0.77$. We may say the same about percentage population disabled (disbP) and percentage vacant units (vacantP) with Spearman's $\rho = 0.73$, as well percentage population disabled (disbP) and percentage over 25 without a high school diploma (noHSP) with Spearman's $\rho = 0.75$.

Relevant code utilized to generate our exploratory data visualizations can be found in Appendix A.2.1 (on page 30).

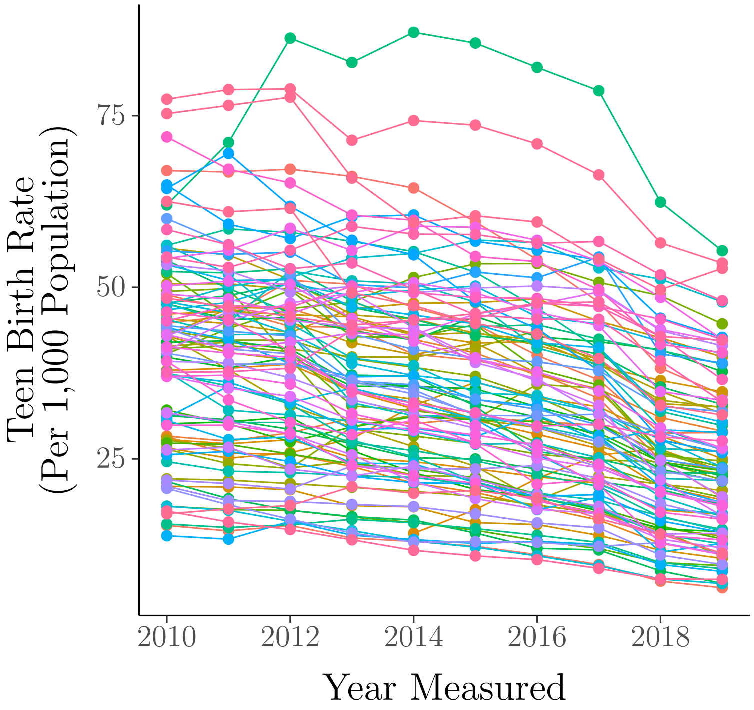
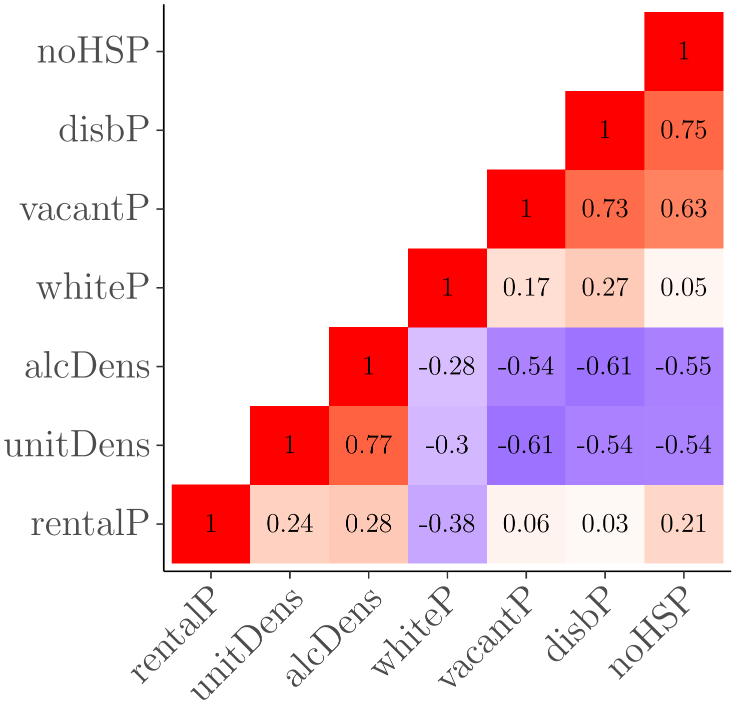


Figure 6.3: Teen birth rate longitudinal trajectories for VA counties.



	rentalP	unitDens	alcDens	whiteP	vacantP	disbP	noHSP
noHSP							1
disbP						1	0.75
vacantP					1	0.73	0.63
whiteP				1	0.17	0.27	0.05
alcDens			1	-0.28	-0.54	-0.61	-0.55
unitDens		1	0.77	-0.3	-0.61	-0.54	-0.54
rentalP	1	0.24	0.28	-0.38	0.06	0.03	0.21

Figure 6.4: Pairwise Spearman rank correlations for all cross-sectional predictors.

6.3 Regression analyses via linear mixed-effects models

We now turn our attention to our regression analyses. We utilized the `nlme` package to run linear mixed-effects models with fixed effects structure defined in the typical specification of regression models and random effects defined to allow a random intercept per county. Specific code utilized in our regression analyses, including syntax for specification of fixed effects and random effects, can be found in Appendix A.2.2 (on page 32).

While our interest is to study the effects of our various predictors (both cross-sectional and longitudinal) on our longitudinal response variable, we have limited data to perform our analyses. As such, we should be concerned with statistical power of our models prior to performing any interpretation. We will consider both extensive (have many terms) and minimal (having few terms) models in our following analyses, beginning from the more extensive models and performing "backwards elimination" to reduce model complexity as we go. While multiple comparisons correction is certainly an important concept to be aware of, there is no consensus as to the appropriate technique to account for them in the context of longitudinal analyses.

We begin with the results from our "fully-specified" model (utilizing all predictors in our dataset) in Table T.3 (on page 19). In this model we detect statistically significant effects of year (or more generally, the passage of time), population (or adjustment factor for the outcome of deaths), and the model intercept at the $p < 0.001$ level. We additionally detect a marginally statistically significant effect of percentage population white at the $p < 0.05$ level. However, this fully-specified model contains both alcDens and unitDens (which were highly correlated). Additionally, our inclusion of alcDens in this model has also shrunk our sample size (as we saw from our descriptives tables) to a total of 213 observations (less than 20 per term including the intercept). As such, we halt our interpretation of this model at this point.

We next examine an almost "fully-specified" model in Table T.4 (on page 20) created by modifying the previous model to simply remove the alcDens term. Our results showcase similar estimates as before, with statistically significant effects detected for year (or time), population (adjustment covariate for our deaths count response), and model intercept at the $p < 0.001$ level. We once again detect a statistically significant effect of percentage population at the $p < 0.05$ level. However, interestingly enough, we now see the emergent detection of a statistically significant effect of percentage population disabled at the $p < 0.05$ level. Our number of total observations has slightly increased to 222, and our number of model terms (including intercept) has decreased from 11 to 10, leaving us with just over 20 observations per model term. While this is a more acceptable situation than previously, we still remain cautious with interpretation of this model's results in search of a model with more appropriate statistical power.

We turn now to our third and final model, our relatively "minimally-specified" model based on terms detected to exhibit statistical significance. We choose to keep our percentage of population white and percentage of population disabled predictors due to their previous detection as exhibiting statistical significance. For this reason as well as more rigid study design reasons, we choose to keep year (or time, to preserve the longitudinal nature of this work), population (as an important covariate for our death counts response variable), and teen birth rate (as an important variable in previous studies and our other non-population time-varying covariate). Results from our regression on this model can be found in Table T.5 (on page 21). We once again detect statistically significant

	Dependent variable:
	Deaths
Year	2.591^{***}
	(0.374)
Population	0.0001^{***}
	(0.00001)
Teen.Birth.Rate	0.055
	(0.137)
whiteP	-0.306^{*}
	(0.134)
noHSP	-0.192
	(0.533)
disbP	1.084
	(0.662)
vacantP	-0.568
	(0.346)
rentalP	0.130
	(0.218)
unitDens	-0.023
	(0.013)
alcDens	155.014
	(89.969)
Constant	$-5{,}197.549^{***}$
	(754.631)
Observations	213
Log Likelihood	-830.619
Akaike Inf. Crit.	1,687.239
Bayesian Inf. Crit.	1,730.246
Note:	*p<0.05; **p<0.01; ***p<0.001

Table T.3: Regression output for the fully-specified model.

	Dependent variable:
	Deaths
Year	2.298***
	(0.351)
Population	0.0001***
	(0.00001)
Teen.Birth.Rate	−0.056
	(0.129)
whiteP	−0.300*
	(0.137)
noHSP	−0.055
	(0.534)
disbP	1.319*
	(0.644)
vacantP	−0.494
	(0.342)
rentalP	0.197
	(0.225)
unitDens	−0.001
	(0.003)
Constant	−4,611.652***
	(709.356)
Observations	222
Log Likelihood	−870.391
Akaike Inf. Crit.	1,764.783
Bayesian Inf. Crit.	1,805.062
Note:	*p<0.05; **p<0.01; ***p<0.001

Table T.4: Regression output for the fully-specified model (all observations).

effects of year (or time), population (covariate of death counts response variable), and the model intercept, all at the $p < 0.001$ level. We additionally reproduce the previously-seen statistically significant effect of percentage disabled population at the $p < 0.05$ level from our previous model. Interestingly, we detect a statistically significant effect of percentage white population at a more conservative $p < 0.01$ level in this model, compared with previously-seen models. This model features 6 terms (including model intercept), resulting in over 35 observations per term (a noticeable improvement from our previous models).

	Dependent variable:
	Deaths
Year	2.324***
	(0.342)
Population	0.0001***
	(0.00001)
Teen.Birth.Rate	−0.029
	(0.120)
whiteP	−0.366**
	(0.120)
disbP	0.886*
	(0.365)
Constant	−4,656.019***
	(693.055)
Observations	222
Log Likelihood	−866.106
Akaike Inf. Crit.	1,748.211
Bayesian Inf. Crit.	1,775.214
Note:	*p<0.05; **p<0.01; ***p<0.001

Table T.5: Regression output for the minimally-specified model.

At this point, we attempt to perform interpretation of model coefficients in terms of predictor contributions to the response variable. Many of the predictors as currently implemented have widely-varying input ranges, some percentages ranging from 0 to 100, an other predictor as a large, whole number population count, and other variable as a large whole number year value. As such, the relative magnitudes of model coefficient estimates are not very amenable to interpretation. For the sake of interpretation, we can examine the sign (positive or negative) of the estimates, but their magnitudes are not very informative without transformation prior to inclusion in our models.

For our statistically-significant effect of year (or time) in the positive direction at the $p < 0.001$ level, we can conclude that the passage of time has typically resulted in the increase of deaths (adjusted for by population) within each county. In other words, the opioid epidemic has gotten worse throughout our longitudinal data. This is a relatively intuitive assumption given the historical context of the opioid epidemic, but we also must remember that the structure of our data requires the use of a time term to enable our longitudinal analyses.

For our statistically-significant effect of population in the positive direction at the $p < 0.001$ level, we can also conclude that counties with larger populations experience more opioid overdose related deaths. This is also a relatively intuitive assumption, however just as for our time term above, this term is a necessary inclusion in our model as it enables us to examine the deaths count response variable in different counties by adjusting for the county population in the same year. (As an additional note, the extremely small magnitude of this term is likely related to the extremely large magnitude of population within a county. As such, if the variable was transformed, this coefficient estimate would likely be more informative.)

For our statistically-significant effect of model intercept in the negative direciton at the $p < 0.001$ level, we typically do not perform interpretation of model intercept value unless our other predictors followed some sort of centering (e.g., mean-centering) procedure. In that case, the intercept would represent the sample expectation for our deaths count response variable. However, in this case, the widely-varying values of our predictor variables makes interpretation of this term uninformative. Just as with the importance of including time and population predictors in our model, we must include an intercept in our model for it to function as we intend, and to enable our further analyses.

For our statistically-significant effect of percentage population white in the negative direction at the $p < 0.01$ level, we can conclude that higher percentages of a county population being white are associated with lower counts of opioid overdose related deaths. Moreover, higher percentages of the population considered nonwhite (100% - percentage population white) is positively associated with an increase in opioid overdose related deaths. This is amenable with our previous exposition of SDoH and minority health generally being in a worse state due to systemic racism and discrimination, among other direct and indirect socioeconomic forces.

For our statistically-significant effect of percentage population disabled in the positive direction at the $p < 0.05$ level, we can conclude that higher percentages of a county population being disabled are associated with higher counts of opioid overdose related deaths. While definitions of disability can be viewed as to some extent being socially constructed, and thus this variable falling under SDoH, we may argue that the picture is more complex. Individuals with disabilities do face social inequity, but arguably they also face quite apparent issues in receiving access to adequate services in the form of health parity (e.g., via medicaid) or potentially struggling with financial independence (e.g., with high expenses and/or incompatibilities with many adequately-paying job functions as currently specified). These issues may certainly have socioeconomic components, but we would like to argue they additionally provide unique issues that may otherwise not fit within the SDoH framework.

For our model term of teen birth rate, a time-varying covariate we implemented in our model with extremely rich CHR data and a model term often discussed in other works in this space, we did not detect a statistically significant association with count of opioid overdose related deaths at the $\alpha = 0.05$ threshold. This was an interesting finding, but issues with our number of observations may mean that our current analyses are simply inconclusive to detect the effect with our given statistical power. This variable was likely the most contextual (or environmental) variable included in our models, but given the opportunity to continue this work with more observations, we would like to include more pertinent variables associated with the risk-environment framework described previously.

7 Conclusions

We developed a combined dataset utilizing cross-sectional OEPS data characterizing the opioid-risk environment, longitudinal CHR data characterizing time-varying trends of health variables, and CDC WONDER opioid overdose related death counts. Our initial exploratory analyses suggested missingness in our data that resulted in a loss of statistical power for regression analyses. While we attempted to perform mixed-effects regression utilizing relevant covariates and predictors, we were unable to interpret many of our model results in-depth due to the wide ranges of input values and lack of centering. For our minimally-specified model, we can conclude that the predictors associated with our opioid overdose related death counts response variable in Virginia counties were percentage population nonwhite (100% - percentage population white), percentage population disabled, time-varying population size, time, and model intercept. For future work, we would recommend finding ways to maximize the sample size (e.g., study all counties in the United States rather than counties only in Virginia) to improve statistical power and trustworthiness in analysis conclusions.

8 Acknowledgments

This material is based upon work supported by the National Science Foundation grant IIS: 1945764 *(EAGER: An Open Data Sharing Platform for Substance Use Disorders)*, and was also supported in part by the NSF grant DGE: 1922598 *(NRT-HDR: Transdisciplinary Graduate Training Program in Data-Driven Adaptive Systems of Brain-Body Interactions).* Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9 Bibliography

Bauer, C., Champagne-Langabeer, T., Bakos-Block, C., Zhang, K., Persse, D., & Langabeer, J. R. (2021). Patterns and risk factors of opioid-suspected EMS overdose in houston metropolitan area, 2015-2019: A bayesian spatiotemporal analysis (N. D. Zaller, Ed.). *PLOS ONE*, *16*(3), e0247050. https://doi.org/10.1371/journal.pone.0247050

Brandt, A. M. (1978). Racism and research: The case of the tuskegee syphilis study. *The Hastings Center Report*, *8*(6), 21–29. http://www.jstor.org/stable/3561468

Centers for Disease Control and Prevention, National Center for Health Statistics. (2020). *Multiple cause of death 1999-2019 on cdc wonder online database*. Retrieved December 15, 2021, from http://wonder.cdc.gov/mcd-icd10.html

DeWeerdt, S. (2019). Tracing the us opioid crisis to its roots. *Nature (London)*, *573*(7773), S10–S12. https://doi.org/10.1038/d41586-019-02686-2

Haley, D. F., & Saitz, R. (2020). The Opioid Epidemic During the COVID-19 Pandemic. *JAMA*, *324*(16), 1615–1617. https://doi.org/10.1001/jama.2020.18543

Heyman, G. M., McVicar, N., & Brownell, H. (2019). Evidence that social-economic factors play an important role in drug overdose deaths. *International Journal of Drug Policy*, *74*, 274–284. https://doi.org/10.1016/j.drugpo.2019.07.026

Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, *113*(16), 4296–4301. https://doi.org/10.1073/pnas.1516047113

Jones, M. R., Viswanath, O., Peck, J., Kaye, A. D., Gill, J. S., & Simopoulos, T. T. (2018). A brief history of the opioid epidemic and strategies for pain medicine. *Pain and Therapy*, *7*(1), 13–21. https://doi.org/10.1007/s40122-018-0097-6

Josephs, L., & Brown, S. E. (2017). The jj way®: Community-based maternity center final evaluation report. *Visionary Vanguard Group Inc.*

Kim, H., & Yang, H. (2020). Statistical analysis of county-level contributing factors to opioid-related overdose deaths in the united states. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 5860–5863. https://doi.org/10.1109/EMBC44109.2020.9176465

Kolak, Lin, Paykin, Menghaney, & Li. (2021). *Geodacenter/opioid-policy-scan: Opioid environment policy scan data warehouse* (comp. software; Version v0.1-beta). Zenodo. https://doi.org/10.5281/zenodo.4747876

Kolak, M.A., Chen, Y.T., Joyce, S., Ellis, K., Defever, K., McLuckie, C., Friedman, S., & Pho, M.T. (2020). Rural risk environments, opioid-related overdose, and infectious diseases: A multidimensional, spatial perspective. *International Journal of Drug Policy*, *85*. https://doi.org/10.1016/j.drugpo.2020.102727

National Institute on Drug Abuse. (2021, March 11). *Opioid overdose crisis*. Retrieved November 16, 2021, from https://www.drugabuse.gov/drug-topics/opioids/opioid-overdose-crisis

NIH HEAL Initiative. (2021, November 2). *Justice community opioid innovation network*. Retrieved December 7, 2021, from https://heal.nih.gov/research/research-to-practice/jcoin

Provine, D. M. (2011). Race and inequality in the war on drugs. *Annual review of law and social science*, *7*(1), 41–60. https://doi.org/10.1146/annurev-lawsocsci-102510-105445

R Core Team. (2021, November 1). *R: A language and environment for statistical computing* (comp. software; Version 4.1.2). Vienna, Austria. https://www.R-project.org/

Remington, P. L., Catlin, B. B., & Gennuso, K. P. (2015). The county health rankings: Rationale and methods. *Population Health Metrics*, *13*(1). https://doi.org/10.1186/s12963-015-0044-2

Rhodes, T. (2002). The 'risk environment': A framework for understanding and reducing drug-related harm. *International Journal of Drug Policy*, *13*(2), 85–94. https://doi.org/https://doi.org/10.1016/S0955-3959(02)00007-5

RStudio Team. (2021). *Rstudio: Integrated development environment for r* (Version 2021.9.1.372). Boston, MA. http://www.rstudio.com/

University of Wisconsin Population Health Institute, & Robert Wood Johnson Foundation. (n.d.). *County health rankings*. Retrieved November 16, 2021, from https://www.countyhealthrankings.org

University of Wisconsin-Madison Social Science Computing Cooperative. (2016). *Mixed models: Testing significance of effects*. Retrieved December 15, 2021, from https://www.ssc.wisc.edu/sscc/pubs/MM/MM_TestEffects.html

U.S. Department of Health and Human Services. (2021, October 27). *About the epidemic*. Retrieved November 16, 2021, from https://www.hhs.gov/opioids/about-the-epidemic/index.html

Wilkinson, R., Marmot, M., for Europe, W. H. O. R. O., Project, W. H. C., for Health, W. I. C., & Society. (2003). *Social determinants of health: The solid facts*. World Health Organization, Regional Office for Europe. https://books.google.com/books?id=QDFzqNZZHLMC

A Appendix of relevant code

This section serves to provide complete code used in this project. Earlier text may reference this section to retain focus on the outcomes of the work rather than on the code. All scripts were created using GNU R version 4.1.2 (R Core Team, 2021) and in RStudio version 2021.9.1.372. (RStudio Team, 2021)

A.1 Data preprocessing scripts

A.1.1 preprocessing_oeps.R

```r
# preprocessing_oeps.R
# Faysal Shaikh
# fshaikh4@gmu.edu
# Nov. 16, 2021
#
# This script serves to preprocess OEPS data files.
# For access to OEPS data see https://geodacenter.github.io/opioid-policy-scan/.
#
library(git2r)
library(readr)
library(dplyr)
library(stringr)

# # establish top of git repository for file path purposes
# repo <- repository('.') # if placed in same repo as listed above
# cwd <- workdir(repo)

# # specify file paths relative to top of git repository
# OEPS_data_path <- file.path(cwd, 'data_final')
# all_files <- list.files(OEPS_data_path)

# TEMPORARY: set cwd as file path of specifically-downloaded OEPS data for project
cwd <- file.path('./OEPS-downloaded-data/OEPS_DOWNLOAD_2021-11-16/data/') # 23 files, less than 92 from above
data_fnames <- list.files(cwd)

# initialize empty lists for file names and variable names to add to later
OEPS_fnames_list <- list()
OEPS_df_list <- list()

# first pass: loop through all data files and create R objects
for (fname in data_fnames) {
  # use fname to name data frame
  name_stem <- str_split(fname, '_')[[1]][1] # pull text before underscore

  OEPS_fnames_list[as.character(name_stem)] = file.path(cwd, fname) # add fnames to empty OEPS_fnames_list created above
  OEPS_df_list

  assign(paste0('OEPS_data_', name_stem), # add data frames to appropriately-named R objects
         read_csv(file.path(cwd, fname),
                  col_types = cols(COUNTYFP = col_integer(), STATEFP = col_integer())
    )  %>% rename_with(toupper, ends_with('ear')) # handle issue with Year vs year vs YEAR
  )

  OEPS_df_list[name_stem] = paste0('OEPS_data_', name_stem) # tie R object names to name_stems in empty OEPS_df_list created above
}

# second pass: loop through all R objects and execute sequential (in order) cumulative pairwise merges
OEPS_data_combined <- get(OEPS_df_list[[1]]) # start with first
# for (df_name in OEPS_df_list[2:length(OEPS_df_list)]) {
for (df_name in OEPS_df_list[2:length(OEPS_df_list)]) {
  OEPS_data_combined <- OEPS_data_combined %>% merge(get(df_name))
}

# drop columns we don't want
OEPS_data_combined <- OEPS_data_combined %>%
  subset(select = -c(YEAR, STATEFP, state, name, note, county))

# save preprocessed data as CSV file
OEPS_data_combined %>% write.csv(file.path(cwd, 'OEPS_data_combined.csv'), row.names = FALSE)
```

A.1.2 preprocessing_chr.R

```
# preprocessing_chr.R
# Faysal Shaikh
# fshaikh4@gmu.edu
# Nov. 15, 2021
#
# This script serves to preprocess CHR data files.
# For more information see https://github.com/fshaikh4/CHR-data-repo
#
library(git2r)
library(readxl)
library(dplyr)

# establish top of git repository for file path purposes
repo <- repository('.')
cwd <- workdir(repo)

# specify file paths relative to top of git repository
CHR_national_data_path <- file.path(cwd, 'data_raw', 'national-data-excel-files')
all_national_files <- list.files(CHR_national_data_path)

# initialize empty lists for file names and variable names to add to later
CHR_fnames_list <- list()
CHR_varnames_list <- list()

```

```r
# first pass: loop through all national data files and create R objects
for (national_file in all_national_files) {
  # pull year and save filenames for each year into empty CHR_fnames_list created above
  year <- substr(national_file, 1, 4) # slice year from filename
  CHR_fnames_list[year] = file.path(CHR_national_data_path, national_file) # add fnames to fname list

  # create separate data frames for each excel file
  assign(paste0('CHR_data_', year), # assign each data frame to 'CHR_data_year' variable
         read_excel(
           file.path(CHR_national_data_path, national_file),
           sheet = 'Ranked Measure Data',
           skip = 1
           ) %>% select(-matches(c('...[0-9]', 'Unreliable'))) # ignore numbered (duplicate) or "unreliable" variables
         )

  # add each variable name to empty CHR_df_list created above
  CHR_varnames_list[year] = paste0('CHR_data_', year)
}

# second pass: loop through created R objects and discover columns common between all years
keep_cols <- CHR_varnames_list[[1]] %>% get() %>% names() # start with first set of variables
for (varname in CHR_varnames_list){
  # each pass, successively remove columns until only those found in all years are left
  keep_cols <- keep_cols %>% intersect(varname %>% get() %>% names())
}

# third pass: loop through objects and only keep the common columns
for (varname in CHR_varnames_list){
  assign(varname, varname %>% get() %>% select(keep_cols) %>%
    # also add year suffix to changing variables (not FIPS, state, county)
    rename_with(~paste(., varname %>% substr(10,14)), -c(FIPS, State, County)) # add year to end
  )
}

# final pass: merge all data by FIPS, State, County
CHR_data_combined <- CHR_varnames_list[[1]] %>% get() # start with first set as combined
for (varname in CHR_varnames_list[2:length(CHR_varnames_list)]) { # since above uses 1st variable, loop range starts from 2nd variable and beyond
  CHR_data_combined <- CHR_data_combined %>% merge(varname %>% get())
}

CHR_varnames_list['combined'] = 'CHR_data_combined' # add to list after loop

# save preprocessed data as CSV file
```

```
CHR_data_combined %>% write.csv(file.path(cwd, ’data_processed’, ’CHR_data_combined.csv’), row.names=FALSE)
```

A.1.3 preprocessing_merge.R

```
# preprocessing_merge.R
# Faysal Shaikh
# fshaikh4@gmu.edu
# Dec. 15, 2021
#
# This script serves to load, merge, and perform any additional
# preprocessing steps for preprocessed CHR and OEPS data files (previously
# handled by preprocessing_chr.R and preprocessing_oeps.R).
#
library(readr)
library(dplyr)
library(reshape2)

# specify working directory structure
cwd <- file.path(’.’)
data_path <- file.path(cwd, ’data’)
code_path <- file.path(cwd, ’code’)
out_path <- file.path(cwd, ’output’)

# load data files into data frame objects
oeps_fname <- file.path(data_path, ’OEPS_data_combined.csv’)
chr_fname <- file.path(data_path, ’CHR_data_combined.csv’)
outcomes_fname <- file.path(data_path, ’outcomes.tsv’)

oeps_df <- read_csv(oeps_fname)
chr_df <- read_csv(chr_fname)
outcomes_df <- read_tsv(outcomes_fname)

# perform preprocessing on individual data frames and merge
## oeps data: create FIPS variable (to merge on)
##   select the following variables:
##     unitDens: Number of housing units per square mile of land area
##     alcDens:  Number of alcohol outlets per square mile
##     rentalP:  Percentage of occupied housing units that are rented
##     vacantP:  Percentage of housing units vacant
##     essnWrkE: Percentage of population employed in Essential Jobs as
##               defined during the COVID-19 pandemic
##     disbP:    Percentage of civilian non-institutionalized population
##               with a disability
##     noHSP:    Percentage of population age 25 years and over with less than
##               high school degree
##     whiteP:   Percentage of pop. with race identified as white alone

oeps_df_preproc <- oeps_df %>% rename(FIPS = COUNTYFP) %>%
  select(FIPS, unitDens, alcDens, rentalP, vacantP,
         essnWrkE, disbP, noHSP, whiteP) %>%
  filter(FIPS>51000 & FIPS<52000)

## chr data: create FIPS variable (to merge on)
##   select the following variables:
##   Teen Birth Rate X: Number of births per 1,000 female population
##                      ages 15-19 during year X.
##
##   [want to select X from 2010 to 2019]

chr_df_preproc <- chr_df %>% mutate(FIPS = as.integer(FIPS)) %>%
  select(c(’FIPS’, contains(’Teen Birth Rate’))) %>%
  select(-contains(c(’2020’, ’2021’))) %>%
  filter(FIPS>51000 & FIPS<52000)

chr_df_preproc_long <- melt(chr_df_preproc,
        id.vars=’FIPS’,
        measure.vars=colnames(chr_df_preproc %>%
                                select(contains(’Teen Birth Rate’))),
        variable.name=’Year’, value.name=’Teen Birth Rate’
)

levels(chr_df_preproc_long$Year) <- c(2010, 2011, 2012, 2013, 2014,
                                      2015, 2016, 2017, 2018, 2019)

chr_df_preproc_long$Year <- chr_df_preproc_long$Year %>%
  as.character() %>% as.numeric()

## outcomes data: create FIPS variable (to merge on)
## remove unnecessary variables

outcomes_df_preproc <- outcomes_df %>% select(-’Notes’) %>%
  rename(FIPS = ‘County Code‘) %>%
  mutate(Deaths = as.numeric(Deaths)) %>%
  mutate(Population = as.numeric(Population)) %>%
  select(County, FIPS, Year, Deaths, Population)

## merge data frames!
combined_df_preproc <- outcomes_df_preproc %>%
  merge(chr_df_preproc_long) %>%
  merge(oeps_df_preproc)
```

A.2 Data analysis scripts

A.2.1 exploratory-data-viz.R

```
# exploratory-data-viz.R
# Faysal Shaikh
# fshaikh4@gmu.edu
# Dec. 16, 2021
#
# This script serves to perform exploratory data visualization on preprocessed
# and combined data following the use of
# preprocessing_merge.R, preprocessing_chr.R, and preprocessing_oeps.R

library(dplyr)
library(ggplot2)
library(tikzDevice)
```

```

# requires existing combined_df_preproc following preprocessing_merge.R

# deaths (unadjusted) by year spaghetti plot
tikz(’output/death-longi-trajectories.tex’, width=4, height=4)
combined_df_preproc %>% ggplot(aes(x=Year, y=Deaths,
                                   group=FIPS, color=factor(FIPS))) +
  geom_line() + geom_point() +
  scale_x_continuous(breaks=seq(2010, 2019, 2)) +
  xlab(’Year Measured’) +
  ylab(’Deaths (Unadjusted)’) +
  theme_classic() + theme(text = element_text(size=12),
                          legend.position=’none’,
                          aspect.ratio=1,
                          axis.title.x = element_text(margin=margin(t=10)),
                          axis.title.y = element_text(margin=margin(r=10)))
dev.off()

# crude death rate (per 1,000 population) by year population spaghetti plot
tikz(’output/crude-death-rate-longi-trajectories.tex’, width=4, height=4)
combined_df_preproc %>% ggplot(aes(x=Year, y=Deaths/Population*1000,
                                   group=FIPS, color=factor(FIPS))) +
  geom_line() + geom_point() +
  scale_x_continuous(breaks=seq(2010, 2019, 2)) +
  xlab(’Year Measured’) +
  ylab(’Crude Death Rate \n (Per 1,000 Population)’) +
  theme_classic() + theme(text = element_text(size=12),
                          legend.position=’none’,
                          aspect.ratio=1,
                          axis.title.x = element_text(margin=margin(t=10)),
                          axis.title.y = element_text(margin=margin(r=10)))
dev.off()

# teen birth rate (per 1,000 population) by year population spaghetti plot
tikz(’output/teen-birth-rate-longi-trajectories.tex’, width=4, height=4)
combined_df_preproc %>% ggplot(aes(x=Year, y=‘Teen Birth Rate‘,
                                   group=FIPS, color=factor(FIPS))) +
  geom_line() + geom_point() +
  scale_x_continuous(breaks=seq(2010, 2019, 2)) +
  xlab(’Year Measured’) +
  ylab(’Teen Birth Rate \n (Per 1,000 Population)’) +
  theme_classic() + theme(text = element_text(size=12),
                          legend.position=’none’,
                          aspect.ratio=1,
                          axis.title.x = element_text(margin=margin(t=10)),
                          axis.title.y = element_text(margin=margin(r=10)))
dev.off()

# rank-based spearman correlation coefficients of all cross-sectional variables
## create cross-sectional variables subset dataframe
cs_df <- combined_df_preproc[,c(1,7:dim(combined_df_preproc)[2])] %>% unique()

## calculate spearman correlations
cormat <- cor(cs_df %>% select(-c(FIPS, essnWrkE)),
              use='complete.obs', method='spearman')

## correlation matrix formatting helper functions
## Get lower triangle of the correlation matrix
get_lower_tri<-function(cormat){
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}
## Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
## reorder correlation matrix
reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}

## perform reformatting of correlation matrix
tri_cormat <- cormat %>% reorder_cormat() %>% get_lower_tri()
melted_cormat <- tri_cormat %>% melt()

## plot spearman correlations on heatmap
tikz('output/cs-spear-corr.tex', width=4, height=4)
melted_cormat %>% tidyr::drop_na() %>%
  ggplot(aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + xlab('') + ylab('') +
  geom_text(aes(Var1, Var2, label = round(value, 2)),
            color = "black", size = 3) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1)) +
  theme_classic() +
  theme(axis.text.x = element_text(angle=45, vjust=1, size=12, hjust=1),
        axis.text.y = element_text(size=12), legend.position='none') +
  coord_fixed()
dev.off()
```

A.2.2 regression-analyses.R

```
# regression-analyses.R
# Faysal Shaikh
# fshaikh4@gmu.edu
# Dec. 16, 2021
#
# This script serves to perform regression analyses via linear mixed-effects
# models on preprocessed and combined data following the use of
# preprocessing_merge.R, preprocessing_chr.R, and preprocessing_oeps.R
#

library(dplyr)
library(table1)
library(kableExtra)
library(printr, quietly=T)
library(nlme)
library(stargazer)

# requires existing combined_df_preproc following preprocessing_merge.R

# generate descriptive statistics table via table1 package
## apply labels & units
label(cs_df$Deaths) <- 'Opioid overdose related deaths'
label(cs_df$Population) <- 'County population'
label(cs_df$`Teen Birth Rate`) <- 'Teen birth rate (adjusted by 1,000)'

label(cs_df$unitDens) <- 'Housing units density (per sq. mi.)'
label(cs_df$alcDens) <- 'Alcohol density (per sq. mi.)'
label(cs_df$rentalP) <- 'Percentage rental units'
label(cs_df$vacantP) <- 'Percentage vacant units'
label(cs_df$noHSP) <- 'Percentage over 25 without HS diploma'
label(cs_df$disbP) <- 'Percentage poulation disabled'
label(cs_df$whiteP) <- 'Percentage population white'

## continuous render helper function
my.render.cont <- function(x) {
  with(stats.apply.rounding(stats.default(x), digits=2),
       c('','Mean (SD)'=sprintf('%s (%s)', MEAN, SD)))
}

## generate tables
### without NA removal
table1(~ unitDens + alcDens + rentalP + vacantP + noHSP + disbP + whiteP,
       render.continuous=my.render.cont,
       data=cs_df) %>%
  t1kable(format='latex') %>% writeLines('output/cs-descriptives.tex')

### with NA removal
table1(~ unitDens + alcDens + rentalP + vacantP + noHSP + disbP + whiteP,
       render.continuous=my.render.cont,
       data=cs_df %>% tidyr::drop_na()) %>%
  t1kable(format='latex') %>% writeLines('output/cs-descriptives-NA-removed.tex')

# specify nlme linear mixed effects model
## change variable name due to issues with nlme
combined_df_preproc <- combined_df_preproc %>%
  rename(Teen.Birth.Rate = `Teen Birth Rate`)

## fully-specified model (< 20 observations per term)
lme(fixed = Deaths ~ Year + Population + Teen.Birth.Rate + whiteP +
      noHSP + disbP + vacantP + rentalP + unitDens + alcDens,
    random = ~ 1|FIPS,
    na.action = na.exclude,
    data=combined_df_preproc
) %>% stargazer(type='latex',
                star.cutoffs = c(0.05, 0.01, 0.001),
```

```r
                out='output/full-model-alcDens.tex'
)

## almost fully-specified model (around 20 observations per term)
lme(fixed = Deaths ~ Year + Population + Teen.Birth.Rate + whiteP +
      noHSP + disbP + vacantP + rentalP + unitDens,
    random = ~ 1|FIPS,
    na.action = na.exclude,
    data=combined_df_preproc
) %>% stargazer(type='latex',
                star.cutoffs = c(0.05, 0.01, 0.001),
                out='output/full-model.tex'
)

## minimal model: only teen birth rate, disabled percentage, and white percentage predictors
lme(fixed = Deaths ~ Year + Population + Teen.Birth.Rate + whiteP + disbP,
    random = ~ 1|FIPS,
    na.action = na.exclude,
    data=combined_df_preproc
) %>% stargazer(type='latex',
                star.cutoffs = c(0.05, 0.01, 0.001),
                out='output/minimal-model.tex'
)
```