Multilevel Matching in Natural Experimental Studies: Application to Stepping up Counties

Niloofar Ramezani[1], Alex Breno[2], Jill Viglione[3], Benjamin Mackey[2], Alison Evans Cuellar[4], April Chase[2], Jennifer Johnson[5], Faye Taxman[2]

[1]Department of Statistics, George Mason University, 4400 University Drive, Fairfax, VA 22030
[2]Center for Advancing Correctional Excellence, Schar School of Policy and Government, George Mason University, 4400 University Drive, Fairfax, VA 22030
[3]Department of Criminal Justice, University of Central Florida, 12805 Pegasus Drive, Orlando, FL 32816
[4]Department of Health Administration and Policy, George Mason University, 4400 University Drive, Fairfax, VA 22030
[5]Division of Public Health, Michigan State University, 200 East 1st St, Flint, MI 48502

Abstract
Among many approaches for selecting match control cases, few methods exist for natural experiments (Li, Zaslavsky & Landrum, 2007), especially when studying clustered or hierarchical data. The lack of randomization of treatment exposure gives importance to using proper statistical procedures that control for individual differences.

In this natural experimental study, which has a hierarchical structure, we plan to evaluate the efforts of 455 counties across the United States to make targeted efforts to improve mental health services and reduce jail utilization over time. Nested within states, counties are clustered on health and social indicators, which affect the likelihood of making improvements in these areas. Similar to a randomized trial, prior to collecting survey data, it is necessary to identify matched control counties as study sites based on an array of state and county covariates. Accounting for the hierarchal structure of data, a blend of various probability-based models are presented to achieve this goal. Methods include multivariable models that control for observed differences among treatment and control groups, shrinkage based LASSO as a variable selection technique, and logistic models.

Key Words: Hierarchical Data, Matching Methods, Survey Study Design, Logistic Models, Stepping Up

1. Introduction

The Stepping Up (SU) initiative is a national effort administered by three national associations (Council of State Government, National Association of Counties, and American Psychiatric Foundation) with the goal to reduce the number of individuals with mental illness in jail. To evaluate the effectiveness of county efforts to reduce jail utilization, the study examines the efforts of 455 SU counties compared to matching non-SU counties. Factors that explain the differences between the SU and non-SU counties should be identified, and used as key matching variables, to ensure that the matched samples are similar in terms of the factors that affect size of the jail population and mental health (MH) resources, such as demographic, health, crime, and other characteristics.

A county grouping methodology developed by CDC was adapted and modified for this study (Givens, Gennuso, Jovaag & Van Dijk, 2017; Kindig, Asada & Booske, 2008). Our contribution was estimating accurate likelihoods of counties being classified as either a SU or a non-SU county. These likelihoods were then used to build scores to find the best matching non-SU county that has similar characteristics to the SU county. Within this paper, the process for identifying quality control counties to their Stepping Up county counterparts is discussed.

1.1 Data

Three primary data sources were used in this study. First, we used the Vera Institute's incarceration trends database to obtain county level incarceration statistics. This data source contained the raw count of pretrial populations and jail populations for each county. These variables, along with population size of each county, calculated the per capita jail and pretrial rates for a county.

Second, we used the County Health Rankings & Roadmaps (CHRR) data, maintained by the University of Wisconsin under funding from the Robert Wood Johnson Foundation, to obtain the health, economic, social, and demographic information for each county. This data source also included data collected from the U.S. Census. Third, we added police and crime data from the Uniform Crime Report.

All 3,100 counties in the United States were included in this study. These counties varied in size and were categorized in three different sizes: small (n=2,882), medium (N=186), and large (N=75). Small counties have a population less than 250,000; medium-sized counties with a population between 250,000 and 750,000; and, large counties with a population greater than 750,000. Small counties had more missing demographic and crime data, and the models performed differently compared to the medium- and large-sized counties. The size of the county is therefore featured in all parts of our analysis.

1.2 Variables

After reviewing several social, health, economic, and demographic variables, and running a variety of descriptive and correlation analysis models, we selected 17 variables to predict the size of the jail (per capita jail population) and county mental health provider rates, as shown in Table 1. The variables were selected to avoid multicollinearity and overfitting.

After running the predictive models including all the following variables, important factors were identified and used in the next phase of the study, which was classification and regression models to choose final key variables for building the matching scores.

Table 1: Description of public health and justice factors used in the models

Variable	*Source*
Demographics of the County	
Size. Indicator variables were created for the three county populations: < 250,000, between 250,000 and 750,000, and over 750,000).	*U.S. Census Population Estimates (2016 in the RWJ County Health Rankings and Roadmaps database)*
Percent of population living in a rural part of the county	*U.S. Census Population Estimates (2016 in the RWJ County Health Rankings and Roadmaps database)*
Median household income	*Small Area Income and Poverty Estimates (2016), in the RWJ database*

Income inequality which reflects the difference between the 80th and 20th income percentile	*American Community Survey, 5-year estimates (2016), in RWJ database*
High school graduation rate	*EDFacts (2015), in the RWJ database*
Percent of population that are African American	*U.S. Census Population Estimates (2016), in RWJ database*
Percent of population that are Hispanic	*U.S. Census Population Estimates (2016), in RWJ database*
Health Care Related Variables	
Number of physically unhealthy days or days an individual indicates they were not feeling well	*Behavioral Risk Factor Surveillance System (2016), in the RWJ database*
Primary care physician rate based on number of physicians in a county	*Area Health Resource File/American Medical Association (2015), in the RWJ database*
Total amount of costs from health care	*Dartmouth Atlas of Health Care (2015), in RWJ database*
Percent of drug treatment services paid by Medicaid	*IMS Institute for Healthcare Informatics (2016), n amfAR Opioid and Health Indicator database*
Indicator of a medical school in the county	*Association of American Medical Colleges (AAMC) list of all U.S. medical school and admission requirements*
Psychiatrists per capita	*American Health Resources File (2019), in RWJ database*
Licensed psychologists per capita, indicating the total number of licensed psychologists divided by the total county population	*American Health Resources File (2019), in RWJ database*
Community MH centers per capita to indicate outpatient services	*American Health Resources File (2019), in RWJ database*
Crime-Related Variables	
Violent crime rate comprised of murder and non-negligent manslaughter, rape, robbery, and aggravated assault	*Uniform Crime Reporting – FBI (2014), in RWJ database*
Police per capita indicating number of police officers divided by the total county population	*Uniform Crime Report (2011)*
Outcome Variables	
Jail population per capita, indicating the average daily number of individuals in a jail divided by the total county population	*Bureau of Justice Statistics (2015), in the VERA Incarceration trends report*
Jail pretrial population per capita, indicating the average daily number of individuals in jail awaiting trial divided by the total county population	*Bureau of Justice Statistics (2015), in the VERA Incarceration Trends Report*
Mental health provider rate, indicating the total number of mental health providers divided by the total county population	*CMS, National Provider Identification (2017), reported in the RWJ database*

2. Methods

Accounting for the hierarchal structure of data, a blend of various probability-based models were used to achieve the goal of identifying quality control counties.
Methods include multivariable models to control for observed differences among treatment (SU) and control (non-SU) groups. Beta regression, random forest as dimension reduction technique, least absolute shrinkage and selection operator (LASSO) as variable selection technique, and logistic models to define variable weights are the statistical models used in the process of building matching scores and finding the best (globally optimized) control county match for each SU treatment county.

2.1 Beta Regression

Practitioners commonly use regression models to analyze data with response variables that seem to be related to other predictor variables. The linear regression model, in particular, is widely used in application. It is not, however, appropriate for situations where the response is restricted to the interval (0, 1) because it may yield fitted values for the variable of interest that exceed the actual lower and upper bounds (Ferrari & Cribari-Neto, 2004). Formally introduced in political science (Paolino, 2001), beta regression model differs from traditional linear regression as it models a dependent variable with values restricted between 0 and 1 following a beta distribution. Beta regression has been shown to perform better than linear regression on a transformed variable for percentage data (Paolino, 2001).

A beta regression model is written similar to a linear regression model but uses a link function, commonly a logit link function, to account for the non-linear relationship that exists between response and predictors. The model is as below,

$$logit(y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where $logit(y)$ is the log-odds of the ratio response variable, β_0 is the intercept, the $\beta_i s$ $(i = 1,2, \dots, k)$ are the slope coefficients of the beta regression model, and the $X_i s$ $(i = 1, 2, \dots, k)$ are the predictors.

In this study, beta regression models were fitted to SU and non-SU counties to identify health, mental health, crime and criminal justice (CJ), socioeconomic, and demographic variables that significantly predict two outcomes: per capita jail population and mental health provider rate. These models illustrated how different significant indicators performed while predicting jail and mental health outcomes within the SU counties compared to the non-SU counties.

SU efforts are focused on decreasing the number of individuals with mental illness in jails; therefore, it is important to identify factors that contribute into predicting per capita jail population as well as mental health provider rate, which is a representation of counties' mental health infrastructure or resources.

2.2 Variable Selection and Dimension Reduction

After initial regression models, next step was to come up with a parsimonious model with fewer variables, yet approximately as informative, to avoid overfitting. To achieve this goal, a set of dimension reduction and variable selection methods were applied to the variables picked from the regression models.

Dimension reduction techniques applied within the classification and regression algorithms and shrinkage-based approaches assist researchers in handling data sets with higher dimensions of variables and observations efficiently within a reasonable timeline without overfitting the statistical models (Ramezani, 2020).

2.2.1 Classification Random Forest

Random forest is a supervised learning technique in data mining, which can be used for prediction and classification of the outcomes and identifying important predictors (Ramezani, 2020). Random forest models can be used for both categorical and continuous variables in two forms of classification random forest and regression random forest, respectively.

Breiman (2001) proposed random forests, which use multiple decision trees in the following manner. First, each tree is constructed using a different bootstrap sample of the data, and next, random forests build the classification or regression trees in a way to build the most optimal group of trees. In other words, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction of the response, while sorting the predictors based on their level of importance in the model (Breiman, 2001). This strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines, and neural networks, and is robust against overfitting of the data (Liaw & Wiener, 2002).

A tree is called a classification tree when the dependent variable is categorical, or binary, and it is called a regression tree when the dependent variable is continuous. When classification decision trees are used while building the random forest for modeling categorical response variables, the respective random forest is referred to as classification random forest.

Modeling whether each county is involved in the SU efforts or not, classification machine learning random forest models were fitted to find the list of important variables and to determine which variables, with their corresponding threshold, explain best the classification of counties to SU or non-SU category.

2.2.2 Shrinkage-Based model

Two extensions of linear regression models for higher dimensions of data are ridge regression and LASSO, which are used for regularization. When applying multiple linear regression models, more features and factors are added to the model compared to the simple linear regression. Having more features may seem like a perfect way for including more information in a model, as well as improving the accuracy of the trained model by reducing the loss within the loss function. This is because the trained model will be more flexible and will take into account more parameters that can potentially explain more variation of the response variable. However, adding more features to the model is not always a good idea due to the increased likelihood of overfitting. Overfitting happens in the presence of many features in regression models. If not filtered and explored up front, some features can be more destructive to the accuracy of the model than helpful by repeating information that are already expressed by other features and adding sample specific noise to the models (Ramezani, 2020). Therefore, to avoid low-quality over fitted and poorly trained models, one of the most common correction mechanisms called regularization is used.

Least absolute shrinkage and selection operator, abbreviated as LASSO, is a linear modeling technique, which also performs regularization on variables in consideration. LASSO is an extension built on regularized linear regression. LASSO method not only punishes high values of the coefficients β, but also actually sets them to zero if they are not relevant and gives a model with fewer features. Therefore, LASSO is considered a variable selection technique too (Kukreja, Löfberg, & Brenner 2006). This feature makes it more useful for this study compared to ridge regression

While identifying the strength and direction in which different predictors contribute to predicting whether a county is SU or not, LASSO was used to select relevant predictors. This step was performed to ensure only variables that are significant in this classification, and do not overlap with the other variables used in matching, are chosen.

2.3 Logistic Regression

Logistic regression is a member of generalized linear model family, which allows one to form a multiple regression relation between a response variable and several predictors or independent variables. Logistic regression is useful for predicting the presence or absence of a characteristic or outcome based on values of a set of predictor variables. Logistic regression takes advantage of using a link function to account for the binary nature of the response variable. Logit link function is the most common link function for binary outcomes within logistic regression (Agresti, 2018).

Quantitatively, the relationship between the occurrence of the event of interest and its dependency on several predictors can be expressed as:

$$p = {}^{1}\!/_{1+e^{-z}},$$

where p is the probability of an event taking place. The probability varies from 0 to 1 on an S-shaped curve and z is the linear combination of predictors including their coefficients. Logistic regression involves fitting an equation of the following form to the data:

$$z = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where β_0 is the intercept of the model, the $\beta_i s$ $(i = 1,2, \ldots, k)$ are the slope coefficients of the logistic regression model, and the $X_i s$ $(i = 1, 2, \ldots, k)$ are the predictors.

In Logistic Regression, probability of the outcome is measured by the odds of occurrence of an event. Change in probability is not constant or linear with constant changes in the values of the predictor variables. This means that the probability of success given the predictor variable (X) has a non-linear logistic function. The most common form of logistic regression uses the logit link function, which results in the following equation:

$$logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Due to the importance of mental health and jail-related outcomes in explaining the differences between SU and non-SU counties, such variables along with a variety of health, economic, social, criminal justice, and demographic variables were included in the fitted logistic regression models. The models were performed to identify factors that differentiated between the SU and non-SU counties, and this allowed for estimating the related likelihoods for the county matching step. The probability (p), which is estimated in the logistic regression models within this study, is the probability of a county being part of the SU effort as opposed to $(1 - p)$, which is the probability of being a non-SU county.

3. Results

First, bivariate correlations and summary statistics were performed to understand the distributions and relationships of the variables in our dataset of all the U.S. counties. There were three advantages to performing these correlation analyses. The correlations amongst the jail population- and metal health resource-related outcome variables allowed us to identify which variables to utilize as outcome variables for the beta regression models. Ultimately, three outcome variables (mental health provider rate, jail per capita rate, and pretrial per capita rate) were selected. Using the per capita rates allowed us to minimize the impact of the population size within each county as well as harmonize variables across different counties. The correlations amongst the independent variables provided insight into collinearity. Finally, the correlations between the outcome variables and the independent variables illustrated the relationship among the variables and how to build the best models.

Second, when different variables in our dataset were fully explored, inferential models were implemented. The first set of models were beta regression models fitted to two county-level outcome variables: per capita jail population and mental health provider rate. Per capita pretrial jail population outcome was also modeled but not included in this paper due to the similarity of the regression model to the model predicting per capita jail population. The per capita jail population consists of the average number of individuals in jail on any given day as a percentage of the population of the county. The mental health provider rate consists of the number of psychologists, counselors, social workers, and psychiatrists in a county that can provide services. This step is important in identifying the variables that affect these outcomes so they can be included in the list of potential key variables that could distinguish between the SU and non-SU counties.

Beta regression models revealed variables that predicted the per capita jail population and mental health provider rate (a measure of mental health resources available in each county). Significant health-related variables were: Number of physically unhealthy days within a given month (30 days), per capita rate of psychiatrists, and percent of drug treatment paid by Medicaid. Significant demographical factors included high school graduation rate and size of the county for medium and large compared to small counties. The only significant crime-related factor was police per capita indicating the concentration of police presence in a county. When predicting the mental health provider rate, statistically significant health-related factors were: Number of physically unhealthy days within a given month (30 days), primary care physician rate, healthcare costs, per capita rate of psychiatrists, and percent of drug treatment paid by Medicaid. Significant demographic factors included high school graduation rate, income inequality, and percentage of the county that is rural. The only significant crime-related factor was police per capita.

The next set of inferential models fitted to the data were machine learning random forests. Results from the classification random forest models suggest that number of physically unhealthy days within a given months, primary care physician rate per capita, health care costs, percent of drug treatment paid by Medicaid, police per capita, and county size, in addition to mental health provider rate and jail population outcome variables, were among the variables that played an important role in classifying a county as SU or non-SU. Therefore, these variables, along with mental health provider rate and jail population per capita, were used to create the matching scores for the global optimization algorithm.
Figure 1 shows list of important variables ranked by classification random forests using mean decrease accuracy and mean decrease Gini indices. Figure 2 shows a decision tree of counties being classified as SU or non-SU within the fitted random forest model.

Finally, shrinkage based LASSO variable selection methods were performed to ensure only variables that are relevant and significant to the classification, and do not overlap with the other variables, were chosen. Besides mental health provider rate, jail population per capita, and pretrial population per capita, some other variables were identified while predicting the binary SU variable. These variables included number of physically unhealthy days within a given month (30 days), primary care physician rate per capita, high school graduation rate, income inequality ratio, health care costs, percent of population that are African American, percent of drug treatment paid by Medicaid, police per capita, per capita rate of licensed psychologists, per capita rate of community MH providers, existence of medical schools in a county, and size of the county.

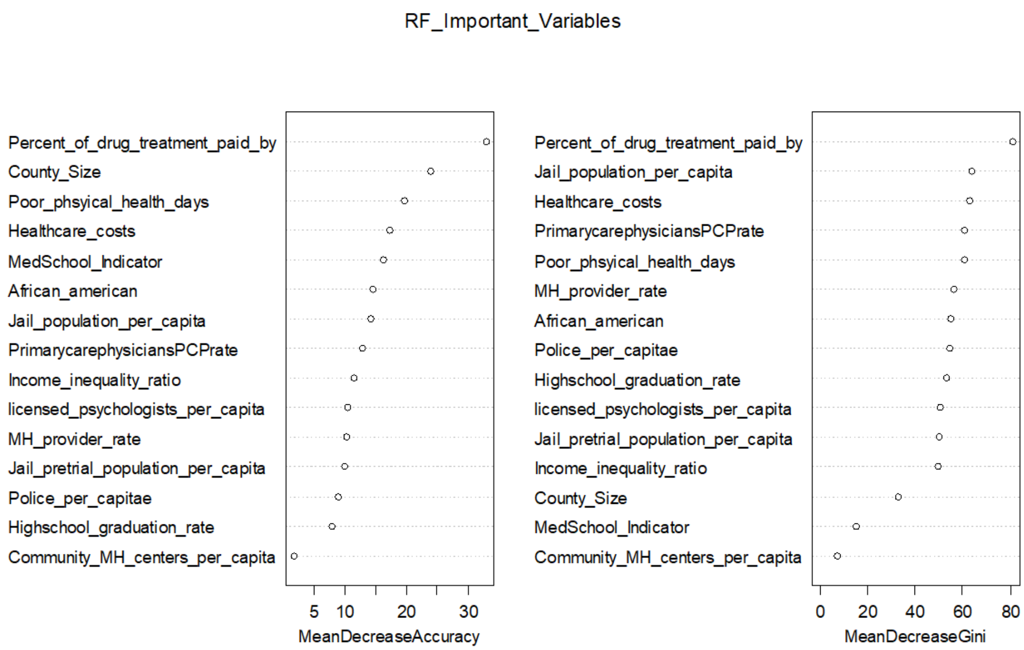

Figure 1: Variable importance plot from classification random forest model

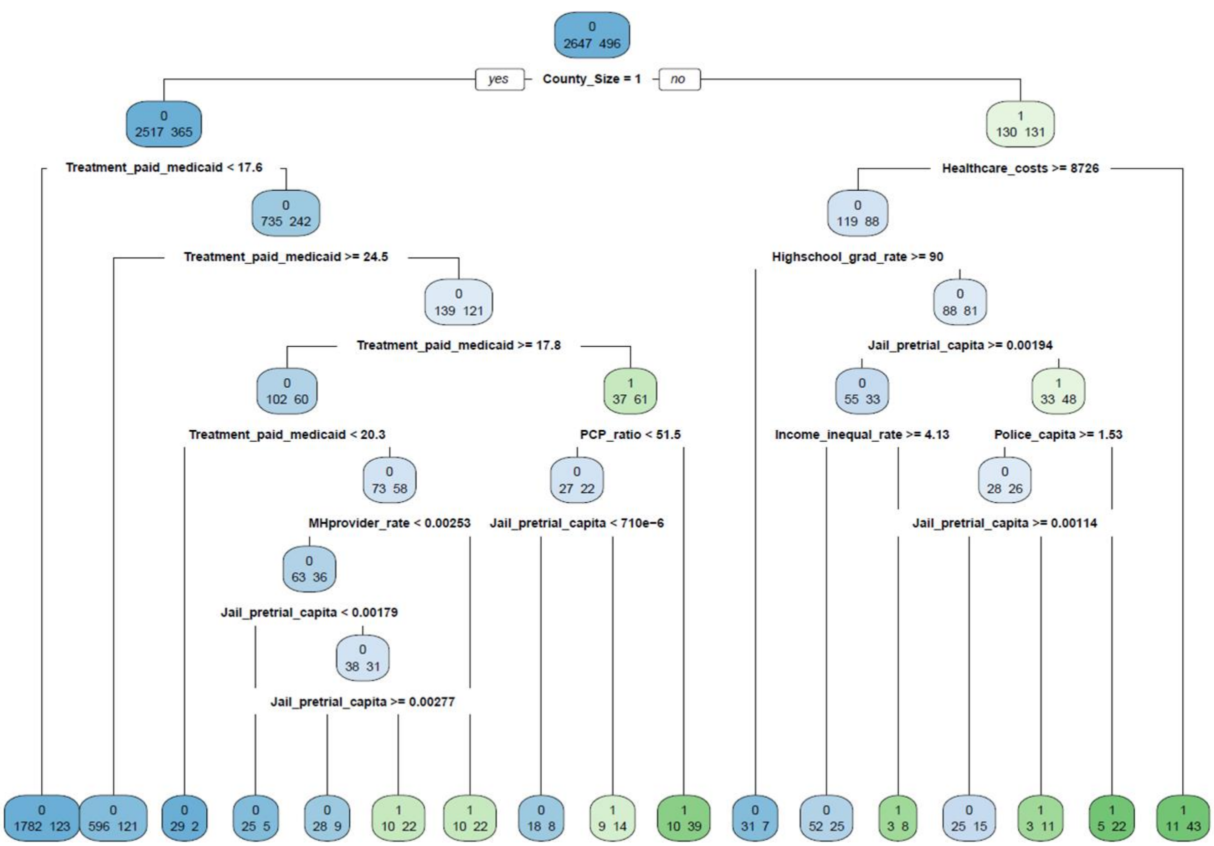

Figure 2: Decision tree of counties classified as SU or non-SU

Third, after finalizing the inferential methods to identify important, relevant and statistically meaningful economic, health, social, and crime/CJ factors that differentiated between SU and non-SU counties, variables that played a key role with regard to this aspect were used within logistic models to estimate the odds of counties classifying as treatment or control counties. This step helped us in calculating the weights that we used in building the likelihood-based matching scores. Table 2 shows the fitted logistic regression using the selected variable, which assisted us in estimating the odds of counties classifying as treatment or control counties.

Logistic regression displayed how different factors interacted with each other when classifying a county as SU or non-SU. Results from our models showed that county level public health factors are key contributors in this classification. Notably, mental health provider rate, number of physically unhealthy days within a given month, health care costs, percent of drug treatment paid by Medicaid, licenses psychologists per capita rate, and medical school indicator variables were among important health related factors in distinguishing between SU and non-SU counties. Jail population and jail pretrial population per capita, and police per capita rate were among important CJ-related variables and finally high school graduation rate and county size were key demographic variables that played an important role in discriminating between SU and non-SU counties across the US.

Table 2. Logistic regression fitted to selected variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.14	0.93	5.529	<0.001
Mental health provider rate	101.911	49.833	2.045	0.041
Jail population per capita	-174.979	40.269	-4.345	<0.001
Jail pretrial population per capita	115.393	54.074	2.134	0.033
Poor physically unhealthy days	-0.468	0.129	-3.622	<0.001
Primary care physician rate	0.003	0.002	1.28	0.2
High school graduation rate	-0.021	0.008	-2.692	0.007
Income inequality rate	0.126	0.114	1.108	0.267
Health care costs	-0.0003	0.00006	-6.301	<0.001
Percent of African American Population	0.009	0.005	1.69	0.09
Percent of drug treatment paid by Medicaid	0.02	0.005	4.098	<0.001
Police per capita	-0.442	0.095	-4.643	<0.001
Licenses psychologists per capita	-764.612	330.662	-2.312	0.021
Community MH centers per capita	-923.235	18898	-0.049	0.961
Medical School Indicator	0.998	0.286	3.493	<0.001
County size Medium vs Small	1.196	0.195	6.148	<0.001
County size Large vs Small	1.866	0.347	5.376	<0.001

Applying a stronger shrinkage to the LASSO models resulted in fewer variables. This step was taken to estimate another set of likelihoods to create the matching scores. Table 3 shows the fitted logistic regression using these selected variables, which assisted in estimating the odds of counties classifying as treatment or control counties.

Table 3. Logistic regression fitted to a more parsimonious list of variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.464	0.55	4.481	<0.001
Mental health provider rate	121.59	39.582	3.072	0.002
Jail population per capita	-84.383	24.296	-3.473	<0.001
Poor physically unhealthy days	-0.259	0.097	-2.657	0.008
Primary care physician rate	0.002	0.002	1.052	0.293
Health care costs	-0.00031	0.00005	-6.288	<0.001
Percent of drug treatment paid by Medicaid	0.016	0.004	3.596	0.0003
Police per capita	-0.355	0.084	-4.201	<0.001
County size Medium vs Small	1.503	0.181	8.32	<0.001
County size Large vs Small	2.644	0.302	8.757	<0.001

This resulted in the following model

$$
\begin{aligned}
logit(\hat{p}) = 2.464 &+ 121.59 MHproviderRate - 84.383 JailPopulationPerCapita \\
&- 0.259 PhysicallyUnhealthyDays \\
&+ 0.002 PrimaryCarePhysicianRate - 0.0003 HealthCareCosts \\
&+ 0.016 TreatmentPaidByMedicaid - 0.355 PolicePerCapita \\
&+ 1.503 Size_{mediumVSsmal} + 2.644 Size_{largeVSsmall}
\end{aligned}
$$

where $\hat{p}$ is the expected probability of a county being a SU county. $logit(p) = \log\frac{p}{(1-p)}$, which is the log of probability of being a SU county compared to being a non-SU county for a given county.

Matching counties. After the logistic regression step was completed, weights were calculated and used in building the likelihood-based matching scores. The matching scores were then used in a matching algorithm to find the closest and best non-SU matched (control) county for each SU (treatment) county. Preference was given to identifying the non-SU counties within the same state to account for the hierarchical nature of the data. The global optimization matching algorithm was written in R (R Core Team, 2017) in a way to choose matching counties for the SU counties within the same state to account for the shared cluster characteristics.

For saturated states, where the number of SU treatment counties was higher than the number of control counties, state's level of involvement in the SU efforts, in addition to other demographic variables were used to first pair the comparable states and then find the best match at the county level.

4. Conclusion and Future Research

Our findings suggest that county-level public health factors are key contributors to not only participating in the SU initiative, but also to broader mental health resources and size of the criminal justice population. Public health factors emerged as more important factors influencing both the size of the county jail population and mental health provider rates compared to crime-related factors. Therefore, they should be considered by researchers in similar studies. Ultimately, more studies need to be performed to answer research questions on how health policies affect crime policy and jail utilization across the U.S.

Some of our results showed that communities that are more economically disadvantaged have fewer mental health resources and tend to have a greater degree of service deserts which contributes to increased use of the jail. There is a need for more studies on the impact of how health policy can undo the effects of mass incarceration, explicitly the degree to which increasing service capacity may address improving mental health functionality in the community and how added community resources can reduce the use of the jail (and incapacitation).

Regarding the statistical methodology for selecting match control cases for natural experiments, more studies are needed to understand different methods to address clustered or hierarchical data. Within this study, we developed an innovative method, which combines an array of statistical and machine learning techniques to come up with estimated likelihood ratios and build best matching scores. Several of the statistical algorithms and machine learning approaches were combined together, such as beta regression, classification random forest and LASSO. These models are not favored by applied researchers possibly due to their complexities and longer run time. The computational burden and time-consuming nature of the existing algorithms for such approaches, based on the current computational capabilities, make such methods less popular among applied researchers and practitioners. This may result in the use of less appropriate models when dealing with non-continuous non-normal response variables. This occurs in both low and high dimensional data analysis. We recommend the use of these methods with a goal to make computationally advanced programming tools more widely available.

Our method of using a globally optimized matching method is another contribution to the field, which made the matching procedure more efficient and accurate. The combination of methods we used should help other researchers who find themselves in need of using similar matching procedures. That is, we are extending our hierarchical likelihood-based matching methods to other researchers who are faced with multi-level matching challenges. The approaches used in this study offer new ways of combining a myriad of statistical methods to address a complex problem. This presents a path forward in new matching methods.

Acknowledgements

The I.M. Stepping Up (Implementation Mechanisms of Stepping Up) study is funded by the National Institute of Mental Health (R01 MH118680, PI Taxman and Johnson). All opinions are those of the research team and not of the funding agency.

References

Agresti, A. (2018). *An introduction to categorical data analysis.* John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Givens, M., Gennuso, K., Jovaag, A., Willems Van Dijk, J. (2017). County Health Rankings & Roadmap. *County Report. Madison, WI: University of Wisconsin's Population Health Institute.* Retrieved from http://www.countyhealthrankings.org/2017-county-health-rankings-key-findingsreport.

Kindig, D.A., Asada, Y., Booske, B. (2008). A Population Health Framework for Setting National and State Health Goals. *JAMA, 299*(17), 2081-2083.

Kukreja, S. L., Löfberg, J., & Brenner, M. J. (2006). A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC proceedings volumes, 39*(1), 814-819.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2007, July). Propensity score analysis with hierarchical data. *In Proceedings of the American Statistical Association Joint Statistical Meetings.*

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 325-346.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Ramezani, N. (2020). Modern Statistical Modeling in Machine Learning and Big Data Analytics: Statistical Models for Continuous and Categorical Variables. In *Handbook of Research on Big Data Clustering and Machine Learning* (pp. 135-151). IGI Global.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics, 31*(7), 799-815.

Van Handel, M. M., Rose, C. E., Hallisey, E. J., Kolling, J. L., Zibbell, J. E., Lewis, B., ... & Iqbal, K. (2016). County-level vulnerability assessment for rapid dissemination of HIV or HCV infections among persons who inject drugs, United States. *Journal of acquired immune deficiency syndromes (1999), 73*(3), 323.